

Fall 12-2013

Gene Regulatory Network Analysis and Web-based Application Development

Yi Yang
University of Southern Mississippi

Follow this and additional works at: <https://aquila.usm.edu/dissertations>



Part of the [Computer Sciences Commons](#)

Recommended Citation

Yang, Yi, "Gene Regulatory Network Analysis and Web-based Application Development" (2013).
Dissertations. 32.
<https://aquila.usm.edu/dissertations/32>

This Dissertation is brought to you for free and open access by The Aquila Digital Community. It has been accepted for inclusion in Dissertations by an authorized administrator of The Aquila Digital Community. For more information, please contact Joshua.Cromwell@usm.edu.

The University of Southern Mississippi

GENE REGULATORY NETWORK ANALYSIS AND
WEB-BASED APPLICATION DEVELOPMENT

by

Yi Yang

Abstract of a Dissertation
Submitted to the Graduate School
of The University of Southern Mississippi
in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy

December 2013

ABSTRACT

GENE REGULATORY NETWORK ANALYSIS AND
WEB-BASED APPLICATION DEVELOPMENT

by Yi Yang

December 2013

Microarray data is a valuable source for gene regulatory network analysis. Using earthworm microarray data analysis as an example, this dissertation demonstrates that a bioinformatics-guided reverse engineering approach can be applied to analyze time-series data to uncover the underlying molecular mechanism. My network reconstruction results reinforce previous findings that certain neurotransmitter pathways are the target of two chemicals - carbaryl and RDX. This study also concludes that perturbations to these pathways by sublethal concentrations of these two chemicals were temporary, and earthworms were capable of fully recovering. Moreover, differential networks (DNs) analysis indicates that many pathways other than those related to synaptic and neuronal activities were altered during the exposure phase.

A novel differential networks (DNs) approach is developed in this dissertation to connect pathway perturbation with toxicity threshold setting from Live Cell Array (LCA) data. Findings from this proof-of-concept study suggest that this DNs approach has a great potential to provide a novel and sensitive tool for threshold setting in chemical risk assessment. In addition, a web-based tool “Web-BLOM” was developed for the reconstruction of gene regulatory networks from time-series gene expression profiles including microarray and LCA data. This tool consists of several modular components: a

database, the gene network reconstruction model and a user interface. The Bayesian Learning and Optimization Model (BLOM), originally implemented in MATLAB, was adopted by Web-BLOM to provide an online reconstruction of large-scale gene regulation networks. Compared to other network reconstruction models, BLOM can infer larger networks with compatible accuracy, identify hub genes and is much more computationally efficient.

COPYRIGHT BY

YI YANG

2013

The University of Southern Mississippi

GENE REGULATORY NETWORK ANALYSIS AND
WEB-BASED APPLICATION DEVELOPMENT

by

Yi Yang

A Dissertation

Submitted to the Graduate School
of The University of Southern Mississippi
in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy

Approved:

Chaoyang Zhang

Director

Dia Ali

Nan Wang

Jonathan Sun

Jiu Ding

Ping Gong

Susan A. Siltanen

Dean of the Graduate School

December 2013

DEDICATION

This dissertation is dedicated to my parents and grandma for their love and support.

ACKNOWLEDGMENTS

I would like to thank Dr. Joe Zhang, Dr. Ping Gong, and all the other committee members, Dr. Nan Wang, Dr. Dia Ali, Dr. Jonathan Sun and Dr. Jiu Ding for their advice and help in my research process. I'm especially grateful for my advisors, Dr. Zhang and Dr. Gong, for their generous help in the past five years. Along my path of research and study at The University of Southern Mississippi, they have spent an enormous amount of time giving me guidance, support and suggestions.

I would also like to thank my lab mates: Si Li, Dr. Haoni Li, Andrew Maxwell, Yan Peng, Dr. Xi Wu, Dr. Peng Li and Dr. Ying Li. Their excellent work inspires me and helps me to learn.

I am also very grateful to my parents and grandma for their everlasting support and generous help.

TABLE OF CONTENTS

| | |
|--|------|
| ABSTRACT..... | ii |
| DEDICATION | iv |
| ACKNOWLEDGMENTS | v |
| LIST OF TABLES | viii |
| LIST OF ILLUSTRATIONS..... | ix |
| CHAPTER | |
| I. INTRODUCTION..... | 1 |
| Motivation | |
| Identification of Differentially Expressed Genes | |
| Reconstruction of Gene Regulatory Networks (GRNs) | |
| Live Cell Array | |
| Web-BLOM | |
| Contributions | |
| Dissertation Organization | |
| II. LITERATURE REVIEW..... | 10 |
| Microarray and Live Cell Array Data | |
| Identification of Differentially Expressed Genes | |
| Gene Regulatory Networks Analysis | |
| Web-based Applications for Biological Network Analysis | |
| III. DIFFERENTIAL RECONSTRUCTED TRANSCRIPTION NETWORKS.. | 25 |
| Background | |
| Live Cell Array Data Set | |
| Data Preprocessing | |
| Identification of Differentially Expressed Genes | |
| Reconstruction of GRNs | |
| Inference of Differential Edges to Build Differential Networks | |
| Functional Annotation and Pathway Mapping of Altered Genes | |
| DE Genes Identified | |
| Pathway Mapping | |
| Reconstructed GRNs of DE Genes | |
| Differential Edges and Differential Networks | |
| Linking Pathway Alteration to Toxicity Threshold | |

| | | |
|-----|--|----|
| IV. | GRN RECONSTRUCTION FROM MICROARRAY DATA..... | 46 |
| | Background | |
| | Earthworm Microarray Data Set | |
| | Data Preprocessing | |
| | Statistical Inference of Differentially Expressed (DE) Genes | |
| | Comparison of DE Gene Identification Algorithms | |
| | Mapping DE or Active Genes to KEGG Pathways | |
| | Bayesian Learning and Optimization Model | |
| | Identification of Pathway Perturbations | |
| | Pathway Perturbations | |
| V. | DEVELOPMENT OF WEB-BLOM..... | 66 |
| | Workflow | |
| | Advantages of BLOM for a Web-Based Application | |
| | System Architecture | |
| | Implementation | |
| | A Case Study of Web-Based Rankprod | |
| | A Case Study of Web-BLOM | |
| VI. | CONCLUSIONS AND FUTURE DIRECTIONS..... | 86 |
| | Conclusions | |
| | Future Directions | |
| | REFERENCES..... | 92 |

LIST OF TABLES

Table

| | | |
|----|---|----|
| 1. | A Summary of Web-Based Gene Regulatory Network Analysis Tools..... | 19 |
| 2. | A Summary of PPI Network Analysis Tools..... | 22 |
| 3. | A Summary of Metabolic Network Analysis Tools | 24 |
| 4. | Percentage of the Strength of Select Edge Over That of the Top Edge in the Reconstructed GRN of High Concentration..... | 39 |
| 5. | The Degree of Pathway Perturbation as Related to the Exposure Concentration of Naphthenic Acids (NAs)..... | 42 |
| 6. | Time in Seconds for Web-BLOM to Return the Results for Two Different Large-Sized Networks..... | 85 |

LIST OF ILLUSTRATIONS

Figure

| | | |
|-----|--|----|
| 1. | A Schematic of a GRN..... | 14 |
| 2. | The Ellipse Layout of Regulatory Networks of E.Coli from Ecocyc Website..... | 18 |
| 3. | Microbial Genome Wide Live Cell Reporter System..... | 29 |
| 4. | A Sub-Network of the Global Regulatory Network Extracted By Pathway Tools Using the Layered Layout..... | 34 |
| 5. | The GRN of the Overlap of the 176 Active Genes and the Gene List of the Ecocyc Global Network..... | 36 |
| 6. | Genes Mapped To the Overall Pathway Map Provided By Ecocyc Database..... | 37 |
| 7. | Histograms of Edge Strength Distribution for 30976 Edges (Gene Connectivity) In Four 176-Node Networks..... | 38 |
| 8. | Differential Networks (DNs) Obtained By Comparing Pair-Wise the Networks Reconstructed for Three Chemical Treatments with Those for the Control Treatment..... | 40 |
| 9. | The KEGG Pathway with the Largest Number of Perturbed Genes – Ribosome Pathway..... | 44 |
| 10. | Expression Profiles of the Earthworm Gene “TA1-181564” Across 31 Time Points in Both Control and RDX..... | 50 |
| 11. | Mapped Genes on GABAergic Synapse Pathway on KEGG..... | 56 |
| 12. | Effect of Carbaryl or RDX Exposure and Recovery on Earthworm MGF Conduction Velocity at 31 Sampling Time Points over the Course of 17 days... | 58 |
| 13. | Number of DE Genes Identified at Each Individual Sampling Time Point and the Sum for The Exposure (E01-13) and Recovery (R01-14) Phases..... | 59 |
| 14. | The Frequency of Differential Expression of Identified Differentially Expressed (DE) Genes across 27 Time Points of Exposure and Recovery Phases..... | 60 |
| 15. | Top 50 KEGG Pathways Enriched with Carbaryl- or RDX-affected Genes..... | 61 |

| | | |
|-----|---|----|
| 16. | Differential Networks of the GABAergic Synapse Pathway Consisting of Differential Edges Inferred from Pair-wise Comparison of Reconstructed GRNs Between the Control And RDX- or Carbaryl-Exposed Earthworms..... | 63 |
| 17. | Workflow of Web-BLOM..... | 67 |
| 18. | Three-tier Architecture of Web-BLOM..... | 70 |
| 19. | Architecture of Cytoscape Web..... | 71 |
| 20. | Workflow of a Servlet in Apache Tomcat..... | 74 |
| 21. | Differently Expressed Genes of Wild-Type Rice on Global Pathway Diagram in Omics Viewer..... | 77 |
| 22. | Differently Expressed Genes of Salt-enduring Rice on Global Pathway Diagram in Omics Viewer..... | 78 |
| 23. | Web-BLOM Page for Uploading Sorted Gene Expression Data..... | 79 |
| 24. | Web-BLOM Page for Selecting Input Parameters..... | 81 |
| 25. | Web-BLOM Result Page that Returns Both the Input Matrix From the Genes and Time Points that the User Selected, and an Output Matrix of Confidence Values..... | 83 |
| 26. | Visualization Using a JavaScript library, Cytoscape Web..... | 84 |

CHAPTER I

INTRODUCTION

Motivation

The amount of genomic information available to researchers is increasing exponentially, e.g. microarray data, protein–protein interactions (PPIs) and metabolic reactions. This means that more effort is required to extract meanings from the data. A scientist can search literature and databases for information about genetic elements and their existing associations with other elements and then use this information to infer new associations with other elements. In addition, the research might involve finding interactions between elements both within and between lists. After forming hypotheses about likely candidates for study, the scientist takes them to the wet lab for validation. This investigation process requires downloading data from different data sources, matching identifiers between data lists such as gene lists, and manipulating lists to match elements (e.g. probe IDs) from one list with elements in other lists. Furthermore, results that contain scores can provide a measure of interaction strength, and they usually require constant string matching, ranking and custom sorting, a process which needs to be automated. A web-based approach enables the integration of these research routines. Meanwhile, for end users, web applications do not require a complex installation like their desktop counterparts or include deploy procedures as required for some open-source packages (Pavlopoulos, Wegener, & Schneider, 2008). Therefore, more and more web applications have become popular in handling biological data as shown in Chapter II.

The limitations of many of the existing computational models to infer gene regulatory network are that they suffer from high computational complexity. Therefore, a

lot of web-based applications based on those models, such as miniTUBA (Xiang, Minter, Bi, Woolf, & He, 2007), which is based on Dynamic Bayesian Networks, either return the results to users by email after computation, or only handle a small network. The Bayesian Learning and Optimization Model (BLOM) is much more computationally efficient than other network reconstruction models such as DBN (Li, 2009), and can be used to reconstruct large-scale networks. Web-based applications require little or no disk space on the client and require no upgrade procedure since all new features are implemented on the server. Web-based tools that integrate large sets of microarrays have already helped biologists reveal novel correlations between responses. For example, the web application Genevestigator helped to uncover strong negative correlation between the expression response to salicylic acid and CO₂ in plants (Laule, Hirsch-Hoffmann, Hruz, Gruissem, & Zimmermann, 2006).

As a maturing genomics technology, microarray has been used successfully in discovering disease- or toxicity-related biomarker genes from gene expression profiling mostly at a single time point. However, like disease inception and progression, organismal response to toxicants is a complicated, dynamic process, whose underlying mechanism may be fully uncovered by capturing temporal changes in molecular interactions within perturbed pathways. Among dozens or hundreds of pathways of an organism, finding the ones of interest requires a universal standard to quantify pathway perturbation. Therefore, a novel approach involving differential networks has been developed in this dissertation to quantify pathway perturbation degrees and to help set thresholds in chemical risk assessment. This approach is applied to Live Cell Array (LCA) data in Chapter III and microarray data in Chapter IV.

Identification of Differentially Expressed Genes

Differentially Expressed (DE) genes are often sought out in genomic studies as they are potential candidates of biomarkers (Huang, 2009). Differential expression studies focus on differences between expression levels of one gene in different samples instead of multiple genes within a sample. Various factors inhibit the comparison of expression levels among different genes. These factors include gene length, protein binding affinity, mRNA degradation rates, and biases introduced by experimental preparation. This dissertation investigates both steady-state microarray data (Chapter V) and time-series data (Chapter III and IV). However, time-series data is the focus of this dissertation. Time-series data has two types of DE genes: (1) active genes that display a differential expression within the same treatment across different time points over the entire course of an experiment (Type I DE genes), and (2) genes that exhibit significant changes in the average expression level across conditions at any given time point (Type II DE genes).

In Chapter II, three statistical methods for detecting DE genes are summarized and compared. In Chapter III, I use a one-sample Gaussian Process (GP) regression method developed by (Kalaitzis & Lawrence, 2011) to identify the first type of DE genes (active genes) from Live Cell Array Data. In Chapter IV, we used two two-sample algorithms developed by (Tai & Speed, 2006) and (Stegle et al. 2010) to identify the second type of DE genes from earthworm microarray data.

Reconstruction of Gene Regulatory Networks (GRNs)

Microarray experiments measure the relative amount of mRNA expressed in different experimental conditions because altered concentrations of a specific sequence of mRNA suggest a homeostatic response of the organism to the experimental conditions. Live cell array experiments, on the other hand, measure the fluorescence level of green fluorescent proteins (GFPs) from living bacteria cells (Melamed, Elad, & Belkin, 2012). Fluorescence levels detected by the sensors of a LCA system directly correlate with the expression level of proteins of interest. Both types of expression profiles can be used to infer regulatory relationships, also known as reverse engineering of regulatory networks. GRN reconstruction provides important insight into answering the two basic motivating questions in biological research: (1) what genes regulate what other genes; (2) what genes or groups of genes regulate a specific phenotype.

Many computational approaches have been proposed to generate predictions corresponding with experimental observations, such as information theory (Steuer, Kurths, Daub, Weise, & Selbig, 2002), Boolean networks (Shmulevich, Dougherty, Kim, & Zhang, 2002), differential equations (de Jong, 2002), and Dynamic Bayesian Networks (Zou & Conzen, 2005). The Bayesian Learning and Optimization Model (BLOM) is a model developed by the Computational Biology and Bioinformatics Laboratory (CBBL) in School of Computing at the University of Southern Mississippi for gene network reconstruction (Li, 2009; Wu et al., 2011). Compared with other GRN reconstruction models, BLOM is much more computationally efficient. Additionally, the output of BLOM provides more information about gene-gene interactions including the types of regulation (inhibition or activation), regulation directions (e.g. gene A regulates gene B

or gene B regulates gene A) and strength of inferred interactions, which can be used as a parameter for ranking and belief management (Doderer, Yoon, & Robbins. 2010).

In Chapter IV, BLOM is used to infer GRN from microarray data of earthworms. Coupled with DE gene selection, visualization techniques and pathway mapping, GRN reconstruction with BLOM is aimed to provide hypotheses and predictions to be tested experimentally, which might suggest new subjects for investigation in web labs that otherwise might not be considered in experimental protocol design.

Live Cell Array

Live Cell Array (LCA) is a new technology that quantitatively measures the real-time gene expression in vivo. It is based on the molecular fusion of a reporter gene to gene promoters from select stress-response regulons. Alterations in biological pathways are a rich resource for setting toxicological threshold, which may be more sensitive and mechanism-informed than traditional toxicity endpoints. A novel approach is developed in this dissertation to connect pathway perturbation with toxicity threshold setting.

My approach consists of 6 steps: time-series gene expression data collection, altered gene identification, GRN reconstruction, differential edge inference, mapping of genes with high differential edges to pathways, and establishment of causal relationships between chemical concentration and perturbed pathways. A one-sample Gaussian Process model was used to identify the genes that exhibited significant profile changes across the entire time course. GRNs of different treatments were reconstructed using BLOM and then compared with each other to infer differential edges/interactions. The differentially expressed genes were then mapped to literature-curated biological pathways in EcoCyc, RegulonDB and KEGG databases. Some of these pathways were perturbed to a degree as

high as 70% even at the lowest exposure concentration, implying that the toxicity threshold for Naphthenic Acids (NAs) could be as low as 10 mg/L. Findings from the results of this study suggest that the differential networks (DNs) approach has a great potential in providing a novel tool for threshold setting in chemical risk assessment.

Web-BLOM

Web-BLOM is a web-based tool developed for reconstruction and analysis of GRNs from time-series gene expression data. The tool consists of several modular components: a database, GRN reconstruction model, and a user interface. The database is used to store user information, the data files uploaded to the server, and the results from the reconstruction model. It also provides several functions for gene expression data analysis. The BLOM model, originally implemented in MATLAB, was adopted by Web-BLOM to provide an online reconstruction of large-scale gene regulatory networks.

From the user interface, users can upload their time-series gene expression data to the server and manage all datasets. If the uploaded files pass an examination of the accepted size and file types, the users can then select a subset of genes. Next, the user interface remotely activates the BLOM code that runs on a dedicated server. After Web-BLOM completes the submitted task, it generates a matrix of confidence values that can be used for ranking the interactions among pairs of the selected genes. If users are interested in prioritizing genes for functional screening, they can use Web-BLOM to return a list of interacting genes ranked by confidence values. Web-BLOM integrates feature selection, network reconstruction, and result analysis in an online network environment and it provides a new efficient and convenient software tool for

reconstruction and analysis of gene regulatory networks. Web-BLOM can be accessed through the following URL: <http://tc1.cs.usm.edu:8080/blom2/>

Contributions

One important goal of GRN reconstruction is to discover novel biomarker genes and pathways. In this dissertation, I have made a number of contributions in identifying DE genes and significant pathways. Based on such information, GRNs of interested pathways were reconstructed using reverse engineering algorithms.

In Chapter III, a novel differential networks (DNs) approach is developed to derive toxicity thresholds based on the perturbation degrees in the reconstructed GRNs. Our approach consists of DE gene identification, GRN reconstruction of the altered genes, differential edge identification, Differential Networks (DNs) construction, pathway mapping and pathway perturbation calculation. Using this approach, I make direct connections between treatment dosage and perturbed pathways.

In Chapter IV, BLOM is used to reconstruct GRNs from microarray data of earthworms treated with two sublethal concentrations of chemicals. Type II DE genes are identified using two different two-sample algorithms and are mapped to pathways on Kyoto Encyclopedia for Genes and Genomes (KEGG) (Ogata et al., 1999). Genes of a neurotransmitter pathway are used to reconstruct GRNs of different conditions and two different stages of the experiment. Our network reconstruction results reinforce previous findings that cholinergic and GABAergic synapse pathways are the target of carbaryl and RDX, respectively. We also conclude that perturbations to these pathways by sublethal concentrations of two chemicals are temporary, and that earthworms are capable of fully

recovering. Moreover, this study indicates that many pathways other than those related to synaptic and neural activities were altered during the 6-day exposure phase.

In Chapter V, Web-BLOM is designed in a three-tier architecture model and implemented in MATLAB and Java. Performance tests were done on different platforms showing that Web-BLOM can return the result for a 200-gene data set to different browsers in less than one minute. MATLAB Java Builder toolbox is employed to package BLOM code into Java archive classes. Rankprod (Hong et al., 2006) is used to identify DE genes from steady-state rice microarray data and mapped to RiceCyc (Jaiswal et al., 2006), a Pathway/Genome Database (PGDB) for *Oryza sativa japonica* (Rice) developed in BioCyc format (Karp et al., 2002).

Dissertation Organization

This dissertation is organized as follows: In Chapter I, the motivation and background of analyzing gene regulatory networks are introduced. Then in Chapter II, web-based applications for analyzing different types of biological networks are reviewed as well as three algorithms to identify differentially expressed genes. These algorithms are later used in Chapter III and Chapter IV.

In Chapter III, a live-cell array data set of the prokaryotic model organism *Escherichia coli* is analyzed to identify differentially expressed genes. Differential edges of these genes in the networks of different dosages are compared in order to establish concentration-pathway perturbation causal relationships.

In Chapter IV, a microarray data set from earthworm toxicity exposure experiment is analyzed using Bayesian Learning and Optimization Model (BLOM) to reconstruct GRNs under different treatments and different stages of the same treatment.

In Chapter V, the implementation of Web-BLOM is detailed, and a case study is presented to demonstrate the web-based data pre-processing and GRN analysis.

The dissertation is concluded with Chapter VI by a summary and some proposed future research directions.

CHAPTER II

LITERATURE REVIEW

Microarray and Live Cell Array Data

Microarray Data

DNA microarray is a collection of microscopic DNA spots attached to a solid surface. Each DNA spot contains picomoles of a specific DNA sequence, which can be a short section of a gene or other DNA element that are used to hybridize a cDNA or an anti-sense RNA sample. Microarray data is a valuable source for gene regulatory network analysis. Microarray technology can be used in many areas such as gene expression profiling, comparative genomic hybridization, chromatin immuno-precipitation on Chip (ChIP), single nucleotide polymorphism (SNP) detection, alternative splicing detection and many more. Using earthworm microarray data analysis, as shown in Chapter IV, this dissertation demonstrates that a bioinformatics-guided reverse engineering approach can be applied to analysis of time-series data to uncover the perturbed pathways by certain chemicals.

The result data set from a microarray experiment first needs to be preprocessed prior to the analysis and interpretation of the results, which includes taking the logarithm of the raw intensity values, flagging bad spots, and handling missing values. Preprocessing is a step that extracts or enhances meaningful data characteristics and prepares the dataset for the application of data analysis methods. Missing values are caused by those data cells that are flagged with physical damages to the array such as bad spots, dust, or scratches. They were omitted from the original data files. Typically less than 1% of spots are flagged. Those spots with aberrantly high or low intensity are also

removed from the data. If missing values are treated as intensity value of zero when calculated, it will certainly affect the accuracy and validity of analysis results. Therefore, methods for imputing missing data are used to minimize the effect of incomplete data sets. The “Data Preprocessing” section in Chapter IV introduces the preprocessing methods used in this dissertation.

LCA Data

Compared with oligonucleotide hybridization-based microarray technology (Ehrenreich, 2006), LCAs avoid complex protocols of pre-treatment and high-cost experimental materials, have less interference, and require less testing time (Elad, Lee, Belkin, & Gu, 2008). It involves the generation of a large number of strains that contain transcriptional fusions with fast-folding fluorescent proteins and monitoring their accumulation under some certain treatments (Aichaoui et al., 2012). An advantage of using bacteria as the organisms for LCAs is the facility by which they can be genetically engineered to respond by a dose-dependent signal to environmental stimuli (Melamed et al., 2012).

Like microarray data, LCA data requires several steps of pre-processing to remove noise. For example, some certain types of bacteria such as *Bacillus subtilis* generate auto-fluorescence in the culture and this could become a background noise to the gene expression data. A software tool based on the discrete Kalman filter was developed by (Aichaoui et al., 2012) to provide standardized treatment to LCA data and generate reports on the quality of the data. If there is no auto-fluorescence ($F_{\text{auto}} = 0$), the promoter activity is directly related to the time derivative of the fluorescence divided by

OD₆₀₀. Otherwise, a correction has to be performed and this can be done with the software BasyLiCA (Aichaoui et al., 2012).

An LCA system was constructed from E. coli K12 strain MG1655, and contained a genome-wide library of modified green fluorescent protein (GFP) expressing promoter reporter vectors (Zaslaver et al., 2006). This genome-wide live cell reporter array has been used to study modes of actions of a wide variety of chemicals (Gou & Gu, 2011; Su et al., 2012). Nevertheless, current MOAs are mostly qualitative and focus on differentially expressed genes in canonical pathways. Although some efforts have been made to identify no-observed transcriptomic effect levels, e.g., Ludwig et al. (2011), little has been done to investigate gene interaction alterations in toxicity pathways (i.e. pathway perturbations) that are often inferred using reverse engineering techniques such as a state-space model with hidden variables (Wu et al., 2011).

Identification of Differentially Expressed Genes

Differentially Expressed Genes

Differentially expressed (DE) genes are often sought in genomic studies as they are considered potential candidates of biomarkers (Abeel, Helleputte, Van de Peer, Dupont, & Saeys, 2010). Microarray and Live Cell Array data provide a snapshot of the dynamic gene expression in living organisms. A traditional way is to calculate the fold change and pick the genes with expression ratio significantly larger from 1 (e.g. 1.5 or 2.0). The t-statistic is another frequently used method for identifying DE genes between two putative conditions. Let $X_{g1}, X_{g2}, \dots, X_{gi}, \dots, X_{gN}$ be a set of observed expression values or a set of averaged expression values of gene g in N conditions where $g = 1, \dots, G$ (number of genes detected on microarrays) and $i = 1, \dots, N$. Suppose that these N

conditions have common effect $+\tau$ or $-\tau$ on the expression levels of gene g (Tan, 2010). It is also supposed that the expression noise for gene g is a random variable across all conditions.

Gaussian Process Regression Models for Time Course Data

There are many methods to identify the genes exhibiting the most significant variation, such as a fixed fold-change cut-off method, t-test, ANOVA, Mann-Whitney test, Z-score, and volcano plot. A Gaussian Process one-sample algorithm (Kalaitzis & Lawrence, 2011) was used in Chapter III. A Gaussian Process two-sample algorithm (GP2S) (Stegle et al. 2010) and a multivariate empirical Bayes algorithm (Tai & Speed, 2006) are adopted in Chapter IV of this dissertation. Both methods are based on the Gaussian Process Regression model, which was established following Gaussian Process ($y|x \sim GP(y; m(x), K_y(x_i, x_j))$) and a Gaussian process is a collection of random variables, any finite number of which has a joint Gaussian distribution. The Lawrence method used in Chapter III applies the Gaussian process model to fit time-series data from microarray and establish a likelihood ratio test to rank differentially-expressed genes. In Chapter IV, two methods are used to rank genes. Time & Speed's method uses Hotelling T^2 statistic to rank genes while Stegle's method computes log likelihood ratio of null hypothesis and alternative hypothesis to rank genes.

Gene Regulatory Networks Analysis

A gene regulatory network is an abstraction of indirect gene-gene interactions. It does not represent the physical interactions of genes like a protein-protein interaction network (PPI) does. In biological databases such as the ones listed in the next section,

GRNs are usually manually curated from the literature on a given organism and represent a distillation of the collective knowledge about a set of related biochemical reactions.

Gene-gene interaction is highly dependent on TFs (Penfold & Wild, 2011). Figure 1 is a schematic of a GRN that consists of four genes, three of which encode for TFs (genes 1, 3, and 4) and one of which encodes for a protein that catalyzes the production of metabolite 2 from metabolite 1. The edges between nodes are individual molecular reactions, protein-protein or protein-mRNA interactions. Through all of the edges, the products of one gene affect those of another. A series of edges indicates a hierarchy of such dependences. Circles to itself correspond to a feedback loop of a gene. It could be either a feedback inhibition or feedback activation.

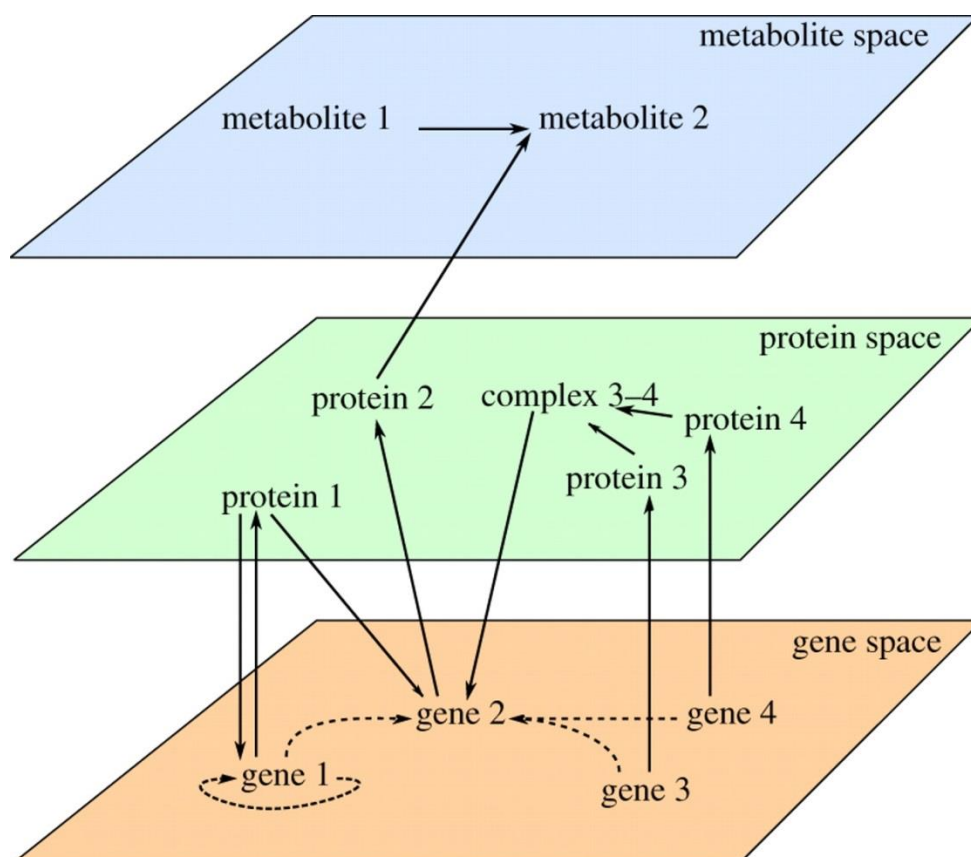


Figure 1. A schematic of a GRN (Brazhnik, de la Fuente, & Mendes, 2002).

One justification of using microarray data or Live Cell Array data to infer regulatory relationships is that the concentration of a transcription factor correlates with the rate at which TF's corresponding mRNA is transcribed (Penfold & Wild, 2011). For instance, higher mRNA concentration of gene 1 (Figure 1) will have a correlated higher concentration of protein 1; the higher or the lower the mRNA concentration of gene 2 will depend on whether the regulation of gene 1 to gene 2 is enhancing or inhibiting. Therefore, many classification and clustering algorithms group genes based on the similarity of expression patterns, and genes within the same group have a higher possibility of regulating each other (Huang, 2009; Kaimal, Bardes, Tabar, Jegga, & Aronow, 2010).

The physical interactions of components at the protein space (green) can be projected onto the gene space (dashed lines) to illustrate the GRN that we try to reconstruct. Therefore, a GRN is an abstraction of the system's chemical dynamics, describing the ways one gene indirectly affects all the genes that it is connected to. This is the reason why most existing GRN inference algorithms suffer low accuracy (Shermin & Orgun, 2009). Various mathematical methods and computational approaches have been proposed to reconstruct GRNs, including Boolean networks, information theory, differential equations and Dynamic Bayesian networks. In the next section, representative web-based applications to analyze or visualize GRNs, PPI networks and metabolic networks are summarized, among which BioCyc, KEGG, RegulonDB, PGDB are used in Chapter III and Chapter IV.

Web-based Applications for Biological Network Analysis

Advantages of the Web-based Approach

Web-based applications greatly facilitate analysis of gene lists and biological networks because they have several advantages over stand-alone tools. They operate through a web browser and are thus easily accessible on any operating system (Murali et al., 2011). Cross-platform capability relieves the application developer from having to worry about a client's configuration. Also because they are accessed through browsers, the user does not need to download and install a copy of the developed software and with databases set up the users can access their uploaded data through different machines. Another advantage of the web approach is that a computationally expensive task can run on a powerful remote server and the user can get access to the results from any personal computer with a browser.

The web approach also saves users the need to install any subsequent upgrades or bug fixes. Web-based applications are also good for collaborative efforts since they allows researchers at different locations to work together handling data or developing models using computational resources that provide more processor-intensive functionality at higher speeds through more robust servers. As a result, through user contribution, web-based applications usually form good repositories for experimental data such as BioCyc (Karp, Riley et al., 2002).

However, web-based applications do suffer from a significant disadvantage in terms of speed of response when dealing with simultaneous user requests and large data transfers between server machines and client sides. Response is significantly slower than stand-alone tools. One possible solution to tackle this problem is to employ Ajax

(Asynchronous JavaScript and XML), which is a combination of several existing web technologies (JJ, 2008). Another disadvantage is that sometimes the users may experience incompatibility when the browser disables some certain element that the web-based application requires, for example, JavaScript and ActiveX.

Web-based Applications for Gene Regulatory Networks

Unlike static graphs of pathways stored in databases such as KEGG or regulonDB (Salgado et al., 2013), both BioCyc and Reactome (Robertson, 2004) provide dynamic visualization of pathway graphs according to user-uploaded gene list. DAVID Bioinformatics tool (Huang et al., 2007) and SideKick (Doderer et al., 2010) allow users to manually add or eliminate genes according to the user's own biological knowledge after the gene list is uploaded.

The customized network display feature facilitates the inspection of regulatory relationships among a smaller set of genes of interest. Take BioCyc as an example. After a gene list is uploaded, the first time the network image of the listed genes is requested, and the drawing is computed at query time via an auto-layout. But, the resulting drawing is cached on BioCyc server so that subsequent queries during the life span of the user session are retrieved from the cache. A user can also select a group of genes of interest for color-highlighting. For example, the user could select the genes with a specific Gene Ontology term (e.g., all genes involved in cell cycle) and display the GRN of this specific biological process. A user can also select the downstream genes that are directly or indirectly regulated by a list of genes. The set of highlighted genes can be redisplayed in a new page, using a layered or ellipse layout as shown in Figure 2.

Some web applications utilize Java applets to implement the same functions described above. For example, PathCase (Elliott et al., 2008) integrates a dynamic (i.e. query-time) visualization applet named GraphViewer that allows users to request changes to a pathway via queries, or to revise the pathway at hand via editing operations, and the system can then visualize the revised pathway or fragment of a network of pathways on the spot.

Both Reactome's SkyPainter (Robertson, 2004) and PathCase perform statistical ranking to find pathways or reactions most related to the user-defined gene list. Both BioCyc and PathCase provide a GO-based gene set enrichment tool and allow flexible setting of p-value for enrichment. The functionalities of a number of other web applications for analyzing GRNs are summarized in Table 1.

Table 1

A Summary of Web-Based Gene Regulatory Network Analysis Tools Surveyed

| Name | Functionality | Key words |
|---|---|--------------------------------------|
| ToppCluster (Kaimal et al., 2010) | Human/mammalian genomes-centered web server application for comparative enrichment and network analysis of multiple gene lists. | GRN; TF |
| BioProfiling.de (Antonov, 2011) | Interpretation of gene or protein lists using enrichment of statistical frameworks. | PPI mapping; |
| ConsensusPathDB (Kamburov et al., 2011) | PPIs, metabolic and signaling reactions and GRNs in a functional association network. | GRN; TF |
| COXPRESdb (Obayashi & Kinoshita, 2011) | Co-expressed gene database for human and mouse and addition of different layers of omics data into the integrated network of genes. | GRN; TF |
| AGRIS(Arabidopsis Gene Regulatory Information Server) (Yilmaz et al., 2011) | Three interlinked databases, AtTFDB, AtcisDB and AtRegNet; predicted and experimentally verified cis-regulatory elements (CREs) and their interactions, respectively. | Promoter regions; GRN; TF |
| DroID (Drosophila Interactions Database) (Murali et al., 2011) | DroID contains genetic interactions and manually-curated PPIs detected from experiments, and predicted protein interactions based on experiments in other species. | PPI mapping; systems integration; TF |
| FANTOM(Functional Annotation Of the Mammalian Genome)(Kawaji et al., 2011) | Database for GRNs of macrophage differentiation. Data comes from cap analysis of gene expression (CAGE), sequencing mRNA 5'-ends with a 2nd-generation sequencer to quantify promoter activities. | Humans; macrophages; mice; TF |
| GeneCAT (Mutwil, Obro, Willats, & Persson, 2008) | Standard coexpression tools such as gene clustering and expression profiling; tools that combine co-expression analysis with BLAST. Arabidopsis & Barley are featured plants. | Coexpression |

Table 1 (continued).

| Name | Functionality | Key words |
|--|---|--|
| GraphWeb (Reimand, Tooming, Peterson, Adler, & Vilo, 2008) | A web server for biological network analysis and module design using a graphical interface. | Cell cycle proteins; humans; PPI mapping |
| HLungDB (Human lung cancer database) (Wang et al., 2010) | Database of lung cancer-related genes, proteins and miRNAs with the experimental evidences through text mining. | Amino acid motifs; genetic epigenesis; humans; |
| MAGIA (miRNA and Genes Integrated Analysis) (Sales et al., 2010) | Integrative analysis of target predictions, miRNA and gene expression data | Gene expression profiling; humans; mRNA; microRNA; |
| mirConnX (Huang, Athanassiou, & Benos, 2011) | A web server for inferring mRNA and microRNA GRNs. mirConnX combines sequence information with gene expression data to create a disease specific, genome-wide regulatory network. | Gene expression profiling; humans; nucleic acid databases; |

Web-based Applications for Protein-protein Interaction Networks

Protein-protein interaction (PPI) networks are usually obtained by two high-throughput experimental techniques. They are yeast two-hybrid screening and mass spectrometry to discover protein complexes (Stark et al., 2011). Most websites for PPI networks act as databases storing manually curated literature evidence instead of web applications. However, some websites incorporate dynamic content generation apart from their databases.

GeneMANIA (Warde-Farley et al., 2010) is a well-developed web application for analyzing PPI networks. For a user-uploaded set of genes or proteins, it matches the gene/protein list with its database which integrates data from many sources, including

physical interactions, pairs of protein-protein interactions, pairs of co-expressed genes, list of proteins with domain similarity, list of proteins in the same pathway or list of proteins located in the same subcellular component (co-localization), and then visualizes the possible molecular associations among the given proteins, thus allowing users to predict functions of uncharacterized proteins on the basis of functions of proteins associated with them. The functionalities of a number of other web applications for analyzing PPI networks are summarized in Table 2.

Table 2

A Summary of Web-Based Protein-Protein Interaction Network Analysis Tools Surveyed

| Name | Functionality | Key words |
|--|--|--|
| AS-ALPS (Shionyu, Yamaguchi, Shinoda, Takahashi, & Go, 2009) | AS-ALPS (Alternative Splicing-induced Alteration of Protein Structure) analyzes the effects of AS on PPI and network through alteration of protein structure. | PPI Database; Protein Structure; Splicing. |
| DIP (Xenarios et al., 2002) | The DIPTM database catalogs experimentally-determined interactions between proteins. | PPI database |
| GeneMANIA (Warde-Farley et al., 2010) | Use GeneMANIA to find new members of a pathway or complex, find additional genes you may have missed in your screen or find new genes with a specific function, such as protein kinases. | Pathway; Factual databases; algorithms. |
| IntAct (Kerrien et al., 2012) | Manually curated database and analysis tools for PPI data. Allow user submissions. | PPI Database |
| MINT (Molecular INTERaction database) (Cesareni, Chatr-aryamontri, Licata, & Ceol, 2008) | MINT focuses on experimentally verified protein-protein interactions mined from the scientific literature by expert curators | PPI database |
| STITCH (Kuhn et al., 2012) | Known and predicted interactions of chemicals and proteins | Protein-chemical Interactions |
| STRING (von Mering et al., 2007) | Known and Predicted Protein-Protein Interactions | PPI database |

Table 2 (continued).

| Name | Functionality | Key words |
|--|---|---|
| GraphWeb (Reimand et al., 2008) | A web server for biological network analysis and module design using a graphical interface. | Cell cycle proteins; humans; PPI mapping. |
| NCG (Network of Cancer Genes) (Syed, D'Antonio, & Ciccarelli, 2010) | Stores data on 736 human genes that are mutated in various types of cancer. For each gene, NCG provides information on duplicability, orthology, evolutionary appearance and topological properties of the encoded protein in a comprehensive version of the human PPI network. | gene expression profiling; humans; mRNA; microRNA; Disease; Mice; TF. |

Web-based Applications for Analysis of Metabolic Networks

Post-genomic research involves an extensive application of high-throughput, non-linear approaches like transcriptomics, proteomics and metabolomics. Among these three "omics" areas, metabolomics is complicated in particular because metabolic networks can consist of multiple types of nodes including enzymes, ion channels, small molecule metabolites and co-factors, making computational prediction difficult. Metscape (Karnovsky et al., 2012) is a web application that generates metabolic networks based on information in KEGG. It predicts a metabolic network with user-uploaded data on enzyme expression levels or compound concentrations. MetaCyc (Karp, Riley, Paley, & Pellegrini-Toole, 2002) is a Pathway/Genome Database (PGDB) hosted on BioCyc website, inheriting BioCyc's rich functionality. BioCyc hosts hundreds of PGDBs related to different diseases of human beings or different organisms, including RiceCyc, which is

used in Chapter V of this dissertation. The functionalities of a few other web applications for analyzing metabolic networks are summarized in Table 3.

Table 3

A Summary of Web-Based Metabolic Network Analysis Tools Surveyed

| Name | Functionality | Key words |
|--|--|--|
| PathExpress (Goffard & Weiller, 2007) | Interpret microarray data by identifying the most relevant metabolic pathways associated with a subset of genes (e.g. DE genes). | Metabolomics |
| ProdoNet (Klein et al., 2008) | A visualization tool for regulatory networks from the PRODORIC bacterial database. It detects common regulators and metabolic pathways from a list of genes or proteins. | Metabolic Network; Network Analysis |
| METANNOGEN (Gille, Hubner, Hoppe, & Holzhutter, 2011) | METANNOGEN stores biochemical reactions needed for the reconstruction of metabolic networks | Metabolic Network; Network Analysis |
| NeAT (Brohée et al., 2008) | A suite of tools for the analysis of metabolic networks, clusters, classes and pathways. | cluster analysis; metabolic networks; protein interaction mapping; |
| KaPPA-View (Tokimatsu et al., 2005) | Overlays gene-to-gene and/or metabolite-to-metabolite relationships as curves on a metabolic pathway map, or on a combination of up to four maps. Pathway maps of KEGG and maps generated from their gene classifications are available. | KEGG; humans; metabolic networks and pathways; metabolome |

CHAPTER III

DIFFERENTIAL RECONSTRUCTED TRANSCRIPTION NETWORKS

Pathway alterations reflected as changes in gene expression regulation and gene interaction can result from cellular exposure to toxicants. Such information is often used to elucidate toxicological modes of action. Alterations in biological pathways are a rich resource for setting toxicant threshold, which may be more sensitive and mechanism-informed than traditional toxicity endpoints. This study developed a differential network (DN) approach to connect pathway perturbation with toxicity threshold setting.

The DNs approach consists of six steps: time-series gene expression data collection, altered gene identification, GRN reconstruction, differential edge inference, mapping of genes with high differential edges to pathways, and establishment of causal relationships between chemical concentration and perturbed pathways. A one-sample Gaussian Process model was used to identify the genes that exhibited significant profile changes across the entire time course. GRNs with respect to different concentrations of chemical treatment were reconstructed using a state-space model and then compared to infer differential edges/interactions which were then mapped to biological pathways in EcoCyc, RegulonDB and KEGG databases. Some of these pathways were perturbed to a degree as high as 70% even at the lowest exposure concentration, implying that the toxicity threshold for Naphthenic Acids (NAs) could be as low as 10 mg/L. Findings from the results in this chapter suggest that our approach has a great potential in providing a novel tool for threshold setting in chemical risk assessment. In future work, the pathway alteration-derived thresholds will be compared with those derived from apical endpoints such as cell growth rate.

Background

Recent advancements in molecular biology technologies, systems biology, and computational toxicology are poised to transform a primarily *in vivo* animal toxicity testing paradigm into a new one dominated by *in vitro* assays (Bhattacharya, Zhang, Carmichael, Boekelheide, & Andersen, 2011; Collins, Gray, & Bucher, 2008; Krewski et al., 2010; National Research Council (U.S.). Committee on Toxicity Testing and Assessment of Environmental Agents, 2007). This new paradigm makes predictions and cross-species extrapolation based on modes or mechanisms of action (MOAs). However, a lot of challenges remain before this transformation becomes a reality, including: how to incorporate toxicity mechanism information into the next generation risk assessment framework, how to obtain quantitative dose-response and time-course data on the perturbed biological processes or pathways, and how to differentiate transient adaptive perturbations from permanent alterations (Bhattacharya et al., 2011; Cote et al., 2012; Edwards & Preston, 2008; Tannenbaum, 2012). Current MOA approaches mostly focus on identifying differentially expressed genes in canonical pathways. Although some efforts have been made to infer non-observable transcriptomic effect levels (e.g., (Ludwig et al., 2011)) or transcriptional benchmark dose values, little has been done to investigate gene interaction alterations in toxicity pathways (i.e. pathway perturbations) that are often inferred from time series gene expression profiling data using reverse engineering techniques such as a state-space model with hidden variables (Li, 2009; Li, Shaw, Yedwabnick, & Chan, 2006; Rangel et al., 2004; Wu et al., 2011).

To address some of the aforementioned challenges, we conducted a proof-of-concept study using a simple and convenient prokaryotic model organism, *Escherichia*

coli, in order to make a direct connection between MOAs and quantitative risk assessment such as toxicity threshold setting (Ben-Israel, Ben-Israel, & Ulitzur, 1998; Currie, 2012). In this study, *E. coli* was exposed to a chemical stressor of three concentrations, and we hypothesized that in stress response: (1) the bacterium had to reassemble biological pathways that differed from their canonical counterparts, and (2) the degree of pathway perturbation was dependent on the exposed concentration. We chose *E. coli* as the test organism also because a microbial live cell reporter array system was constructed recently from its K12 strain MG1655 (Zaslaver et al., 2006). This system contained a genome-wide library of modified green fluorescent protein (GFP) expressing promoter reporter vectors. Live Cell Array (LCA) is a novel technology that enables the acquisition of high-resolution time-series profiles of bacterial gene expression by measuring the fluorescence level in living cells carrying fused fluorescent protein (Elad et al., 2008; Melamed et al., 2012). This genome-wide *E. coli* LCA was used to study MOAs of a wide variety of chemicals (Gou & Gu, 2011; Su et al., 2012) and to collect time-course gene expression data. Zhang et al., (2011) provided the data to our lab for reconstructing differential networks in this dissertation. Here, we report a novel differential networks (DNs) approach we developed to derive toxicity threshold based on the degree of perturbations in reconstructed GRNs. Our approach consists of the following six steps: (1) collect time-series gene expression data of test organisms that received different treatments, (2) identify significantly changed genes involved in normal cellular growth and stress response from the gene expression dataset, (3) reconstruct GRNs of the altered genes under the control and perturbed/treated conditions using reverse engineering techniques, (4) infer differential edges, i.e., interactions gained or lost

from the control to the treated, to construct DNs, (5) annotate and map the genes in the DNs to biological pathways and functions, and (6) establish causal relationships between concentration and pathway perturbation. Using this approach we made direct connections between treatment dosage and perturbed pathways.

Live Cell Array Data Set

LCA is a new technology that quantitatively measures the real-time gene expression. It is based on the molecular fusion of a reporting gene system to gene promoters from select stress response regulons. Compared with oligonucleotide hybridization-based microarray technology (Ehrenreich, 2006), LCAs avoid complex protocols of pre-treatment and high-cost experimental materials, have less interference, and require less testing time (Elad et al., 2008). It involves the generation of a large number of strains that contain transcriptional fusions with fast-folding Green Fluorescent Proteins (GFPs) and monitoring their accumulation under some certain treatments (Aichaoui et al., 2012). An advantage of using bacteria as the organisms for LCAs is that they can be genetically engineered to respond by a dose-dependent signal to environmental stimuli (Melamed et al., 2012). The promoter activity profiles of 96 individual GFP fusions can be obtained at a very high resolution in a microtiter plate format by determining the difference in fluorescence levels at successive time points after the chemical is administered. Promoter activation or suppression can be easily detected by an increase or a decrease in the fluorescence accumulation rate. Figure 3 demonstrates a basic workflow of a typical LCA system.

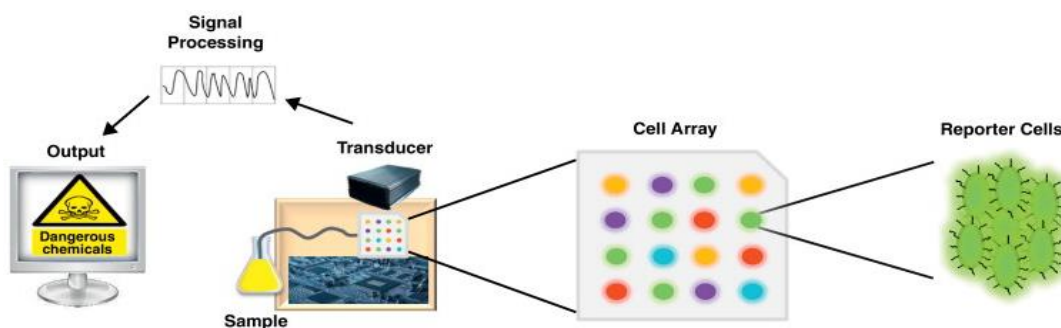


Figure 3. Microbial genome wide live cell reporter system (Melamed et al., 2012).

A time-series data set of dynamic gene expression profiling used in this study was collected in a previous study (Zhang et al., 2011) using the genome-wide *E. coli* LCA made up of twenty-one 96-well plates. Among these 2016 wells, 1870 wells were occupied by 1820 GFP strains with promoter genes (some genes having replicates), another 40 wells were filled with strains with two promoterless genes, and the remaining 106 wells were empty. The standard deviations of the expression values of two promoterless genes were later used to correct for background noise in data normalization steps (Zhang et al., 2011). There was one empty well on each of the first 20 plates and 86 empty wells on the last plate. These empty wells were used to set the cut-offs in the active gene selection step (see below section “Identification of differentially expressed genes”). Optical density (OD) values were measured before treatment. Then the *E. coli* cells received four treatments of a technical mixture of naphthenic acids (NAs, Sigma Aldrich, St. Louis, MO, USA), i.e., 0 (control), 10, 100, or 1000 mg/L of NAs. The fluorescence levels of all 2016 wells were measured every 10 min for three hours, resulting in a dataset of 18 time points. The entire experiment was performed once without any repeat.

Data Preprocessing

The direct estimation of promoter activities from the time-series profile of the fluorescence level change contains high levels of noise (Aichaoui et al., 2012). Therefore, a series of pre-processing procedures need to be done to remove noise. First, raw GFP readings were divided by the OD values measured before treatment. OD values reflected the population density of *E. coli* cells in a well before initiation of a treatment. Because the number of cells in each well might be different due to cell growth, by dividing GFP with OD, we got a preliminary value that reflected the activity of our target genes. Then, the result matrix was smoothed by calculating the moving average of every neighboring three time points. A possible low level of auto-fluorescence of *E. coli* might bring some background noise. To eliminate this type of background noise, the GFP expression produced by the eight promoterless plasmid values were averaged (two promoterless plasmids at four treatments) and subtracted from the values of each gene at the corresponding time point in both experimental and control tests.

Because the promoter activity of each gene might be different at the onset of the experiment, the values of the same gene at time point one in four treatments were averaged, and the differences between the averages and each of the 4 values were calculated. Then, the differences were subtracted from the values of each gene at all of the subsequent time points to eliminate the internal measurement noise. In order to filter the system noise, any value was set to zero if it was less than twice the amount of the standard deviation of the aforementioned processed promoterless values.

If $(\text{GFP}-\text{GFP}(\text{promoter-less average})) > 2 \times \text{STDEV}$ then

$$\text{GFP}_{\text{gene}} = (\text{GFP}-\text{GFP}(\text{promoter-less average}));$$

If $(\text{GFP}-\text{GFP}(\text{promoter-less average})) \leq 2 \times \text{STDEV}$ then

$$\text{GFP}_{\text{gene}} = 0;$$

Where STDEV is the standard deviation of GFP readings of the background (cells with 2 promotorless plasmids).

Identification of Differentially Expressed Genes

DE genes are often sought in genomic studies as they are potential candidates of biomarkers. Different from studies where gene expression is measured at a single time point, time-series experiments have two types of DE genes: Type I: active genes that display a differential expression within the same treatment across different time points over the entire course of an experiment, and Type II: DE genes whose expression vary significantly between different treatments at any given time point. In this study, we identified the first type of DE genes (active genes) using a one-sample Gaussian Process (GP) regression method developed by (Kalaitzis & Lawrence, 2011). In the GP regression model, the continuous trajectory estimation of a gene expression was treated as an interpolation problem on functions of one dimension, given the observations (gene expression time-series). Subsequently, the differential expression of the gene's profile was assigned a marginal log-likelihood ratio by which it was ranked. In the ranking list, the wells with no *E. coli* cells (empty wells) served as cut-off points. Those genes that ranked higher than the highest ranked empty wells in any of the four treatments were included in the active gene list.

The second type of DE genes were identified in the previous study (Zhang et al., 2011) by applying a linear regression model and a cutoff of 1.5-fold change in gene expression at one or more time points between the control and at least one of the three concentrations. These two types of DE genes were pooled together to form the final list of DE genes.

Reconstruction of GRNs

The Bayesian Learning and Optimization Model (BLOM) is used to reconstruct a network of interactions between the identified DE genes for each of the four treatments. BLOM is based on the state space model with hidden variables and an expectation-maximization algorithm to estimate model parameters (Li, 2009; Wu et al., 2011). Pre-processed expression profiles of the identified DE genes were used as the input for BLOM. Like other reverse engineering models such as Dynamic Bayesian Network (DBN) (Zou & Conzen, 2005) and Probabilistic Boolean Network (PBN) (Dorigo, 1994), the outcome of BLOM-reconstructed networks is an $N \times N$ matrix with N being the number of nodes/genes. Each entry of the matrix represents an edge (interaction) between two genes. The connectivity is expressed as confidence level in the form of direction (inward, outward and self-to-self), action type (stimulatory if a positive confidence level value, or inhibitive if a negative confidence level value) and strength (absolute confidence value). The reconstructed networks were visualized using Cytoscape v.2.8.3 (Smoot, Ono, Ruscheinski, Wang, & Ideker, 2011).

Inference of Differential Edges to Build Differential Networks

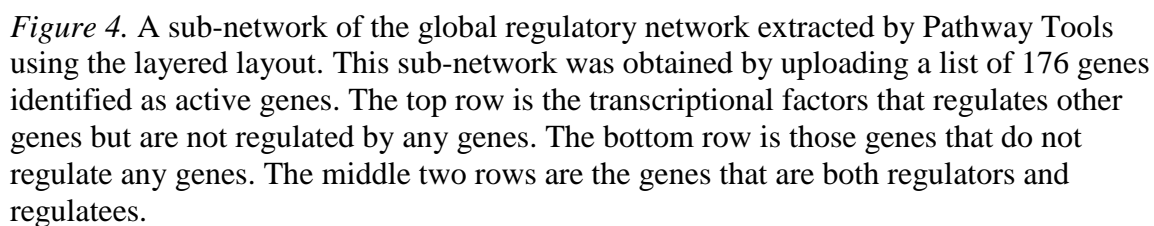
The reconstructed GRNs of all three chemical treatments (low, mid and high concentrations of NAs) were compared pair-wise with that of the control to derive

differential edges, i.e., edges lost or gained from the control to the chemically treated. From the comparison, the following statistics were obtained for each DE gene: total number of edges in each of the four networks, number of gained or lost edges in the treatment networks, and the percentage of edges changed in the treatment networks. Lost edges are those present in the control network but absent in the low, mid, or high concentration network, whereas gained edges are those absent in the control network but present in the chemical treatment network. The following formula was used to calculate the percentage of edges changed as a result of chemical exposure: $(\text{number of gained edges} + \text{number of lost edges}) / (\text{total number of edges in both the control and the exposure networks})$. The changed edges (lost or gained) of all involved DE genes were used to construct differential networks for the three chemical treatments.

Functional Annotation and Pathway Mapping of Altered Genes

Gene Ontology (GO) terms provide information on molecular function, biological processes and cellular component of a gene product. One gene may have multiple GO terms associated with it. The GO tool at www.ecogene.org was used to assign GO terms to the genes of interest (e.g., altered genes). For pathway mapping, we searched the EcoCyc (Karp, Riley et al., 2002), KEGG pathway (Altman, Travers, Kothari, Caspi, & Karp, 2013) and RegulonDB (Salgado et al., 2013) databases. The Pathway Tools software v.15.5 (Karp, Paley, & Romero, 2002) was used to extract pathway mapping information from the Ecocyc database.

The EcoCyc database is one of the two tier-1 databases of BioCyc project (Paley et al., 2012). It stores literature-based curation of approximately 4490 *E. coli* genes and of *E. coli* transcriptional regulation, transporters, and over 360 metabolic pathways (Paley et



The above sub-network was later employed as a benchmark network with which the reconstructed GRNs by BLOM are compared. The reason why there are only 43

genes in the extracted network is that many of our 176 active genes have records in EcoCyc but do not exist in the global GRN in Pathway Tools, which means the regulation relationships of these remaining genes are still waiting to be uncovered.

DE Genes Identified

Each gene was assigned a log likelihood score by the Gaussian Process one-sample algorithm by which the genes are ranked. The 106 wells with no *E. coli* cells (empty wells) in the ranked list were used to set the cut-off level at the highest ranked empty well. As a result, 47, 11, 45, and 101 genes were found with assigned scores higher than the cut-off point and therefore identified as active genes (Type I DE genes) under the control, low, mid, and high concentrations of NAs exposure, respectively, - a total of 111 unique genes. These 111 genes were further pooled together with a group of 85 genes identified by applying linear regression filtering and larger than a 1.5 fold change of the values in the experimental condition versus the values of the same gene in the control condition (Zhang et al., 2011). This resulted in a final list of 176 unique DE genes, with 20 genes appearing in both Type I and Type II DE gene lists.

Among our 176 active genes, 128 genes were found in the 3655-gene global network in the EcoCyc database (Paley et al., 2012), among which 43 genes are non-standalone genes which have 58 edges with each other. This GRN is visualized in Cytoscape as shown in Figure 5. From their annotation information, two genes, GadW and GadX, which were significantly altered in all 3 treatments, are related to acid resistance. The change of their expression potentially triggered the acid response mechanism of the bacteria after their exposure to NAs.

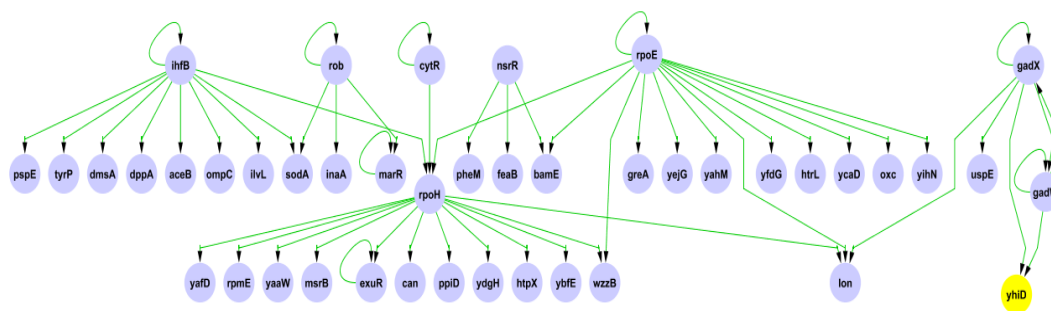


Figure 5. The GRN of the overlap of the 176 active genes and the gene list of the EcoCyc global network.

Pathway Mapping

As of June 2013, 995 genes out of 4490 *E. coli* genes have pathway information in EcoCyc (Paley et al., 2012). Therefore, only 37 out of our 176 genes were successfully mapped to the global pathway map as shown in Figure 6. Because a single gene can appear in multiple pathways, 78 pathways were mapped. Each node in the map represents a metabolite which participates in the reaction and each edge represents the enzyme that catalyzes this reaction. Therefore, most of our genes were mapped onto the edges because the protein products of these genes act as enzymes of the mapped metabolic reactions. The protein products of some of our genes were mapped to nodes because they act as reactants in metabolic reactions. Those genes that are located on the border of the map participate in the trans-membrane activities. The right side of the map shows standalone reactions that do not belong to any pathways and 19 of the 176 active genes that belong to no pathway were mapped to them.

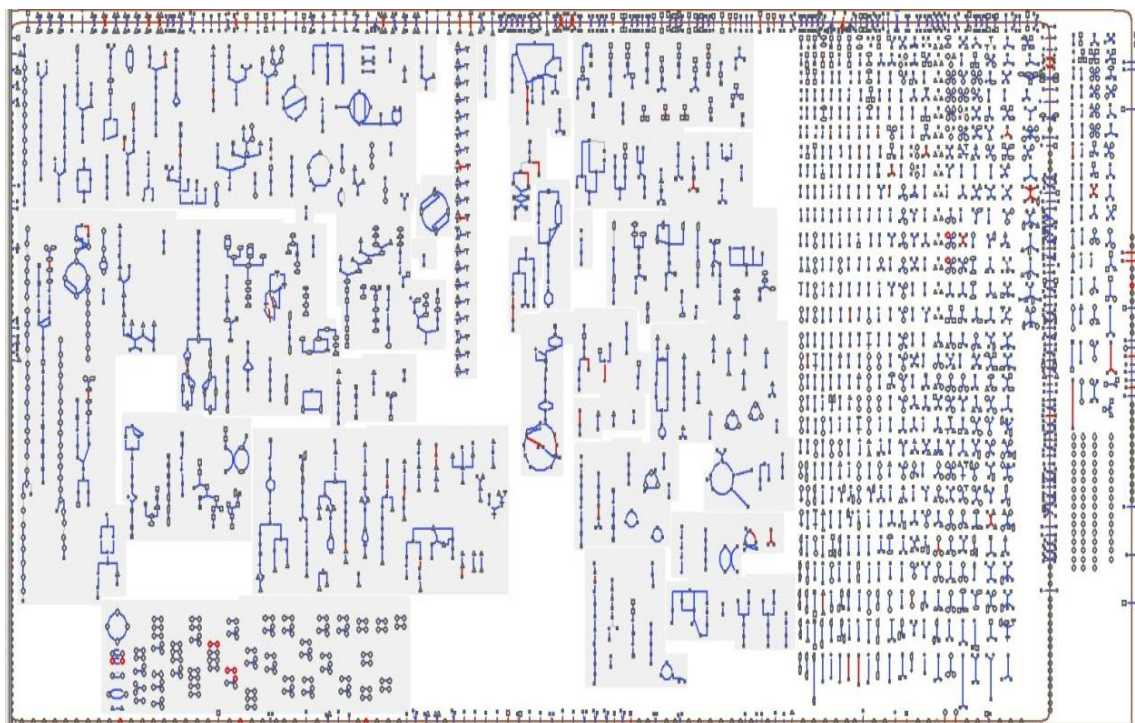


Figure 6. Genes that are mapped to the overall pathway map provided by EcoCyc database. Each node is a reactant and each edge is an enzyme. The shaded chains on the left of the map are metabolic pathways while the rest chains are standalone biochemical reactions. A total of thirty-seven genes among the 176 active genes are mapped to pathways while another nineteen genes among the 176 active genes are mapped to standalone reactions.

Reconstructed GRNs of DE Genes

Four GRNs of 176 nodes (DE genes) were reconstructed using BLOM, one for each treatment. Each network had 30976 (176×176) edges, which were ranked by their strength (i.e., absolute value of confidence level). Obviously, the ~31K edges are not equally important and should not be treated equally. The higher an edge ranks, the more likely it actually exists.

In selecting edges for further analysis, a cut-off level can be set for either the total number of top edges per network or the lowest edge strength allowable in a network. The latest release of EcoCyc database (v. 17.1 as of June 2013) curated 2232 reactions catalyzed by 1500 enzymes that are encoded by 4509 *E. coli* transcription units (genes),

suggesting that the average number of interactions per gene might be very low. We have also observed that the total number of edges in a real-world GRN (e.g., KEGG pathways) generally does not exceed four times the total number of its nodes (genes). Furthermore, we plotted four histograms to show edge strength against edge rank (Figure 7).

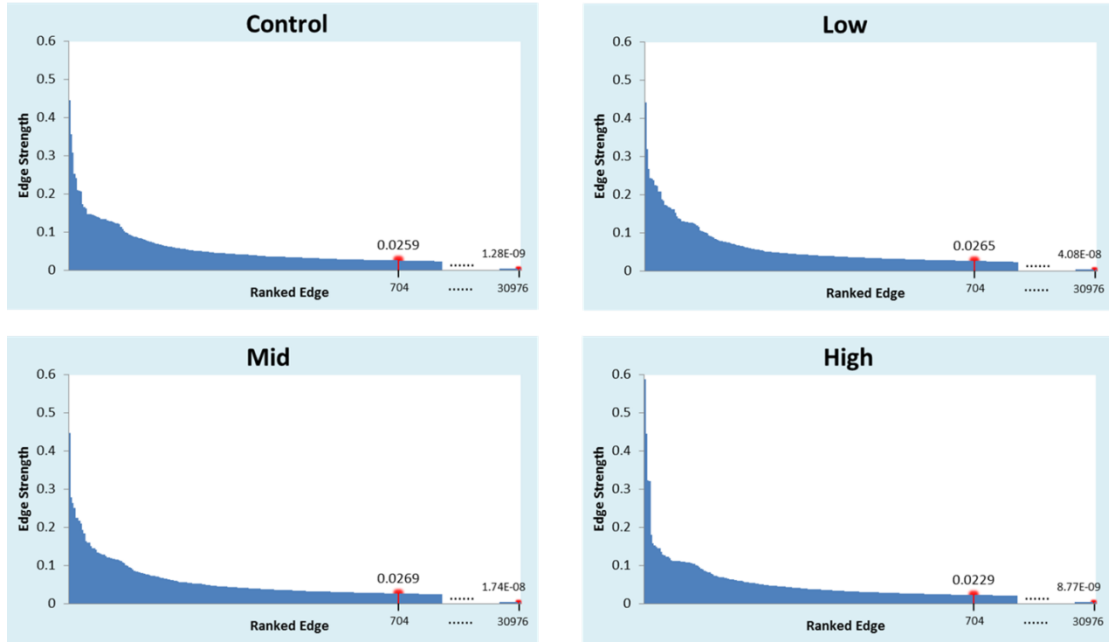


Figure 7. Histograms of edge strength (absolute values of BLOM-inferred confidence level) distribution for 30976 edges (gene connectivity) in four 176-node networks. All edges in each network are sorted by their strength and shown on the X-axis in a descending order. Also shown is the strength of the 704th and the lowest ranked edges.

In the four reconstructed networks, the top ranked 704 edges (4×176 , or 2% of all possible edges) accounted for ~30% of the total strength of all ~31K edges, and the edge strength declined by 94% from the first ranked edge to the 704th edge (Table 4). Therefore, we selected the top 704 edges per network for further differential edges inference and differential network construction.

Table 4

Percentage of the Strength of Select Edge over That of the Top Edge in the Reconstructed GRN of High Concentration

| | Edge strength | Percentage of select edges over top edge |
|-------------|---------------|--|
| Top edge | 0.5878 | 100% |
| 704th edge | 0.0228 | 3.9% |
| 2000th edge | 0.0109 | 1.9% |
| 6000th edge | 0.0042 | 0.7% |

Differential Edges and Differential Networks

As a result of GRN reconstruction in BLOM, both the four 704-edge reconstructed networks and the three differential networks are presented in Figure 8. The four reconstructed networks have 96 (control), 87 (low), 82 (mid), and 99 (high) interconnected nodes/genes, with a total of 117 non-redundant DE genes appearing in these networks. The differential networks were made up of differential edges, i.e., lost and gained edges from the control to a chemical treatment. The number of lost or gained edges were 246 (35% of 704 edges), 299 (42%), and 365 (52%) for the low, mid and high concentration networks, respectively, suggesting a dose-dependence for differential edges. By applying an arbitrary cut-off of 4 differential edges per gene in any one of the differential networks, we removed 37 additional genes and kept the 80 remaining genes for further downstream analysis.

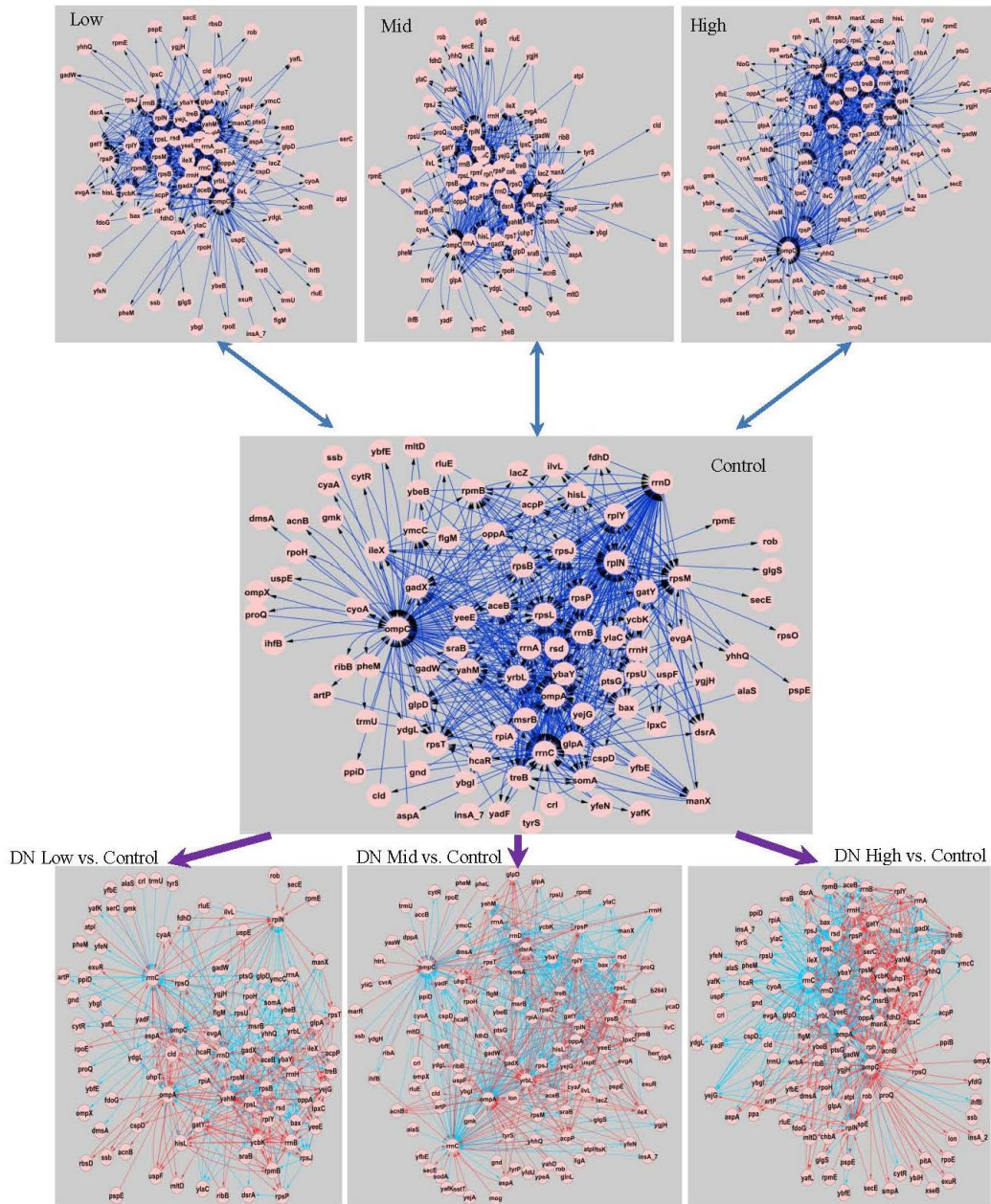


Figure 8. Differential networks (DNs) obtained by comparing pair-wise the networks reconstructed for three chemical treatments with those for the control treatment.

Each of the four reconstructed networks contains 704 edges. In the DNs, red lines represent gained edges (edges absent in the control network but present in the chemical treatment network), whereas blue lines represent lost edges (edges present in the control network but absent in the chemical treatment network).

Linking Pathway Alteration to Toxicity Threshold

The 80 genes possessing a significant number of differential edges were mapped to biological pathways curated in KEGG and EcoCyc databases as well as to GO terms. All but 38 genes were mapped to 35 KEGG pathways. To link pathway alterations to toxicity thresholds, we defined the pathway perturbation degree as the average percentage of edge change per gene for all DE genes involved in any particular KEGG pathway at each exposure concentration (Table 5).

The limited number of concentrations that *E. coli* cells were exposed to and the lack of independent treatment replications in the current study prevented a statistical approach to deriving toxicity thresholds from pathway perturbation degrees. Therefore, for purposes of proof of concept, we turned to establish a simplified causal relationship between concentration and pathway perturbation, which was defined as the perturbation degree at the high concentration being higher than that at both the mid and the low concentrations. Twenty-two perturbed pathways met this definition (Table 5).

Table 5

The degree of pathway perturbation as related to the exposure concentration of naphthenic acids (NAs). Twenty-two KEGG pathways were identified as being altered in a concentration-dependent manner by exposure to NAs (low = 10 mg/l, mid = 100 mg/l, high = 1000 mg/l). The perturbation degree was defined as the average percentage edge change per gene for all genes involved in a particular pathway, and concentration-dependence was defined as high > mid and high > low in perturbation degree. The perturbation degree is expressed in decimals instead of percentages.

| KEGG pathway name (entry) | low | mid | high | Involved genes (total number) |
|---|------|------|------|---|
| Ribosome (eco03010) | 0.32 | 0.40 | 0.45 | rplN, rplY, rpmB, rpsB, rpsJ, rpsL, rpsM, rpsO, rpsP, rpsT, rpsU, rrnA, rrnB, rrnC, rrnD, rrnH (16) |
| Metabolic pathways (eco01100) | 0.45 | 0.42 | 0.69 | aceB, acnB, aspA, gatY, ilvC, lpxC, manX, ribB, rpiA, serC (10) |
| Microbial metabolism in diverse environments (eco01120) | 0.59 | 0.45 | 0.78 | aceB, acnB, manX, rpiA, serC (5) |
| Biosynthesis of amino acids (eco01230) | 0.55 | 0.38 | 0.95 | acnB, ilvC, rpiA, serC (4) |
| Biosynthesis of secondary metabolites (eco01110) | 0.55 | 0.63 | 0.95 | acnB, ilvC, rpiA, yfbE (4) |
| Amino sugar and nucleotide sugar metabolism (eco00520) | 0.61 | 0.52 | 0.79 | manX, ptsG, yfbE (3) |
| 2-Oxocarboxylic acid metabolism (eco01210) | 0.10 | 0.25 | 1.00 | acnB, ilvC (2) |
| Aminoacyl-tRNA biosynthesis (eco00970) | 0.71 | 0.67 | 1.00 | ileX, tyrS (2) |
| Glycerophospholipid metabolism (eco00564) | 0.39 | 0.37 | 0.76 | glpA, glpD (2) |
| Glyoxylate and dicarboxylate metabolism (eco00630) | 0.30 | 0.49 | 0.79 | aceB, acnB (2) |
| Nitrogen metabolism (eco00910) | 0.68 | 0.73 | 0.80 | aspA, yadF (2) |
| Phosphotransferase system (PTS) (eco02060) | 0.41 | 0.37 | 0.62 | ptsG, treB (2) |
| ABC transporters (eco02010) | 0.48 | 0.50 | 0.67 | oppA (1) |

Table 5 (continued).

| KEGG pathway name (entry) | low | mid | high | Involved genes (total number) |
|--|------|------|------|-------------------------------|
| Citrate cycle (TCA cycle) (eco00020) | 0.20 | 0.50 | 1.00 | acnB (1) |
| Fructose and mannose metabolism (eco00051) | 0.33 | 0.24 | 0.52 | manX (1) |
| Glycolysis / Gluconeogenesis (eco00010) | 0.50 | 0.33 | 0.83 | ptsG (1) |
| Lipopolysaccharide biosynthesis (eco00540) | 0.17 | 0.38 | 0.73 | lpxC (1) |
| Oxidative phosphorylation (eco00190) | 0.00 | 0.00 | 1.00 | ppa (1) |
| Pantothenate and CoA biosynthesis (eco00770) | 0.00 | 0.00 | 1.00 | ilvC (1) |
| Propanoate metabolism (eco00640) | 0.20 | 0.50 | 1.00 | acnB (1) |
| Pyruvate metabolism (eco00620) | 0.41 | 0.49 | 0.59 | aceB (1) |
| Valine, leucine & isoleucine biosynthesis (eco00290) | 0.00 | 0.00 | 1.00 | ilvC (1) |

These pathways varied substantially in the number of identified DE genes, from one gene in pyruvate metabolism pathway to 16 genes in ribosome pathway (Figure 9). The perturbation degree varied from 0 (i.e., no edge change of the DE genes at all, such as oxidative phosphorylation at both low and mid concentrations) to 1 (i.e., all edges of the DE genes have changed, such as propanoate metabolism at the highest concentration).

Figure 9. The KEGG pathway with the largest number of perturbed genes– Ribosome pathway. Sixteen genes were perturbed (marked in red) in this pathway. Perturbation degree was determined as the average percentage edge change per gene for all genes involved in a particular pathway as shown in Table 5. (Source from http://www.genome.jp/kegg-bin/show_pathway?ko03010).

Even at the low NAs concentration, some of the pathways were altered to a degree of 50% to 70%, including biosynthesis of amino acids, secondary metabolites, and aminoacyl-tRNA, as well as nitrogen, amino sugar and nucleotide sugar metabolism (Table 5).

While compensatory responses in metabolism or biosynthesis may occur at low chemical concentrations, this suggests that pathway perturbations can be a sensitive endpoint for toxicity if additional evidence links the perturbed pathways to adverse outcomes at the physiological, organismal or population level. A more refined toxicity threshold could be derived using regression approaches such as a benchmark dose method if more concentrations were tested in addition to more replications per treatment.

CHAPTER IV

GRN RECONSTRUCTION FROM MICROARRAY DATA

The etiology of chemically-induced neurotoxicity like seizures is currently not very well understood. Using reversible neurotoxicity induced by two neurotoxicants (carbaryl and RDX) as an example, this study applies the DN approach introduced in Chapter III to analyze time-series microarray gene expression data and uncover the underlying molecular mechanism. The DN approach used in this chapter is more complex because earthworm, unlike *E.coli*, is not a model organism and therefore could not be mapped directly to KEGG pathways. RefNetBuilder was thus employed to map earthworm genes to their counterparts on reference KEGG pathways (Li, Gong, Perkins, Zhang, & Wang, 2011). The results from this study reinforce previous findings that cholinergic and GABAergic synapse pathways are the target of carbaryl and RDX, respectively. We also conclude that perturbations to these pathways by sublethal concentrations of RDX and carbaryl were temporary, and earthworms showed certain restoration of regulation relationships in the GRNs of the 7-day recovery phase. In addition, our study indicates that many pathways other than those related to synaptic and neuronal activities were altered during the 6-day exposure phase.

Background

As a maturing genomics technology, microarray has been used successfully in discovering disease- or toxicity-related biomarker genes from gene expression profiling mostly at a single time point. However, like disease inception and progression, organismal response to toxicants is a complicated, dynamic process, whose underlying mechanism may be fully uncovered by capturing temporal changes in molecular

interactions within perturbed pathways. Using a case study as an example, this chapter demonstrates that pathway perturbations can be inferred using reverse engineering techniques from time-series gene expression data.

In the case study, a microarray gene expression data set was collected at 31 time points from earthworms (*Eisenia fetida*) which received three different treatments (control, RDX - an explosives compound named hexahydro-1,3,5-trinitro-1,3,5-triazine, and carbaryl - a carbamate pesticide) and such previous studies have shown that exposure to sublethal concentrations of RDX or carbaryl led to reversible neurotoxicity in the earthworm (Gong, Inouye, & Perkins, 2007). The objective of the current study is to identify the mechanism of chemical-induced reversible neurotoxicity through reconstruction of perturbed KEGG (Kyoto Encyclopedia of Genes and Genomes) pathways.

Earthworm Microarray Data Set

A species-specific microarray developed for *Eisenia fetida* (Gong, Pirooznia, Guan, & Perkins, 2010) was used. For the convenience of data processing and result analysis, some details of this experiment are introduced in this section. This oligo array contains 43,803 non-redundant 60-mer probes. A synchronized earthworm culture (starting from cocoons) was created and mature worms bearing clitellum and weighing 0.4~0.6 g were chosen for this experiment. Each worm was transferred from artificial soil-based bedding (culture) and housed in an individual glass vial (115 mL in volume). These worms were exposed to carbaryl (20 ng/cm²), RDX (2 µg/cm²) or acetone (solvent control, evaporated overnight) on moistened filter paper lined up inside the vial. These chemical concentrations were selected because they did not cause lethality in preliminary

tests. The entire experiment was divided into three phases: acclimation (4 days), exposure (6 days), and recovery (7 days). The acclimation (A) phase was necessary for the worms to adapt from soil culture to filter paper, and four samplings were taken to establish the “background” baseline under the control condition. Worms were sampled at 13 and 14 time points for all three treatments (control, RDX and carbaryl) during the exposure (E) phase and the recovery (R) phase, respectively.

Sampled worms were measured for conduction velocity of the medial giant nerve fiber (MGF) before being sacrificed by snap-freezing in liquid nitrogen. All yet-to-be-sampled worms were transferred to new vials at the beginning of the next phase. For instance, at the end of exposure phase, all remaining worms were transferred from exposure vials (containing spiked filter paper) to recovery vials (containing non-spiked clean filter paper). Sampled worms were fixed in RNAlater-ICE to preserve RNA integrity at -80 °C.

Total RNA were extracted from at least 5 worms per time point per treatment. RNA samples were hybridized to the custom-designed 44K-oligo array (one sample per array) using Agilent’s one-color Low RNA Input Linear Amplification Kit. The array design was submitted as GPL16366 in Gene Expression Omnibus (GEO). After hybridization and scanning, gene expression data was acquired using Agilent’s Feature Extraction Software (v.9.1.3). A total of 437 good quality arrays were used.

Data Preprocessing

The result data set from a microarray experiment needs to be preprocessed prior to the analysis and interpretation of the results, which includes taking the logarithm of the raw intensity values, flagging bad spots, and handling missing values. Preprocessing is a

step that extracts or enhances meaningful data characteristics and prepares the dataset for the application of data analysis methods. In this study, the following data pre-treatment steps were applied prior to further statistical and computational analyses: (1) feature filtering: flag out spots with signal intensity outside the linear range as well as non-uniform spots; (2) conversion: convert signal intensity into relative RNA concentration based on the linear standard curve of spike-in RNAs; (3) normalization: normalize the relative RNA concentration to the median value on each array; and (4) gene filtering: filter out genes appearing in less than 50% of arrays (i.e., present on at least 219 arrays). There were more than 43,000 genes remaining after this procedure. Figure 10 shows the expression profile and the 95% confidence intervals of an example gene after preprocessing.

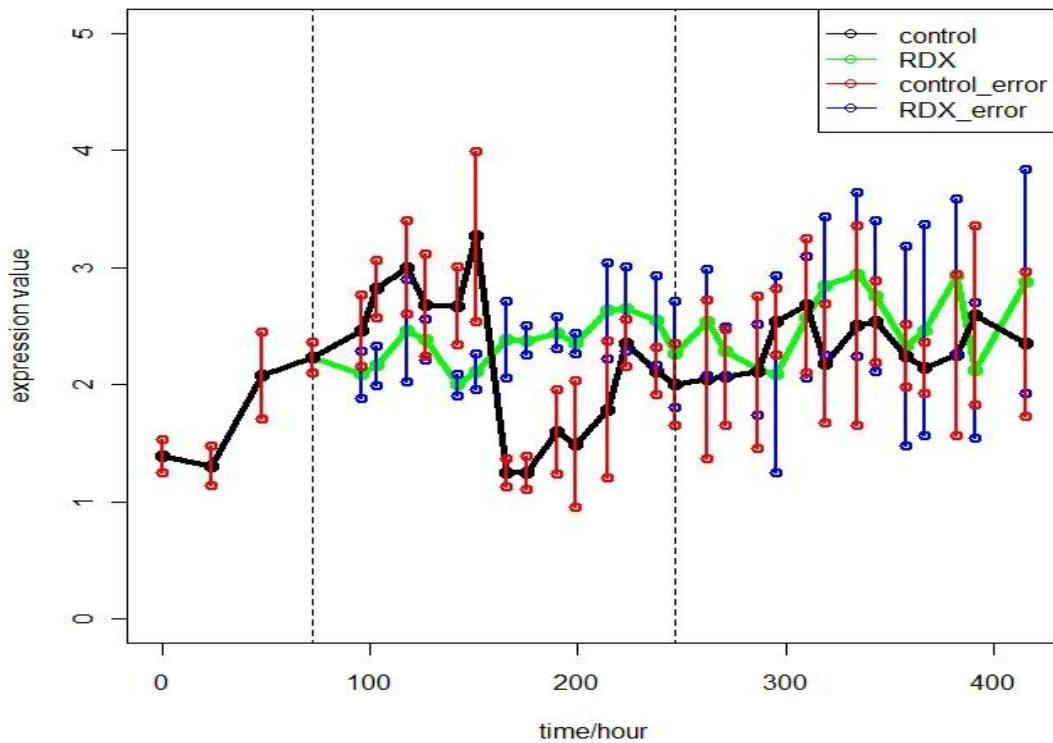


Figure 10. Expression profiles of the earthworm gene “TA1-181564” across 31 time points in both control (black) and RDX (green). The vertical bar stands for the confidence interval of the expression values at the corresponding time point.

Statistical Inference of Differentially Expressed Genes

DE genes at each time point were identified using the “Class Comparison Between Groups of Arrays Tool” in BRB-ArrayTools v.3.8 software package. The collated earthworm array data set was imported without any further normalization or transformation. The tool runs a random variance version of the t-test separately for each gene. It performs random permutations of the class labels and computes the proportion of the random permutations that give as many genes significant at the level set by the user as found in comparing the true class labels. The following two comparisons were conducted to infer genes differentially expressed in response to carbaryl or RDX: controls vs. carbaryl and controls vs. RDX. The following settings were employed: a univariate test

random variance model, multivariate permutation tests with 10,000 random permutations, a confidence level of false discovery rate assessment at 90%, and a maximum allowed number of false-positive genes = 10.

Active genes that exhibited altered longitudinal expression profiles by chemical treatment over the time course were ranked using two algorithms, a multivariate empirical Bayes model implemented in the R software package named “Timecourse” (Tai & Speed, 2006) and a Gaussian Process based two-sample (GP2S) test (Stegle et al., 2010). Because this algorithm requires a universal number of replicates and the earthworm data set contains different number of replicates at different time points, a certain number of replicates were removed to keep 5 replicates across all time points. After correlation of replicates were calculated, the replicate or replicates with the lowest correlation coefficient were removed from the data set. The collated earthworm array dataset was used as the input for both Timecourse and GP2S algorithms.

A cut-off rank equal to twice the number of DE genes was set (Windram et al., 2012). The top ranked genes above the cut-off rank by the two algorithms was intersected. Then the intersection gene list was combined with the DE genes generating the final gene list for downstream analysis.

Comparison of DE Gene Identification Algorithms

Both Stegle’s GP2S algorithm and Lawrence’s method (used in Chapter III) are based on the Gaussian Process Regression model. This model was established following Gaussian Process ($y|x \sim GP(y; m(x), K_y(x_i, x_j))$) and a Gaussian process is a collection of random variables, any finite number of which has a joint Gaussian distribution. The Lawrence method used the Gaussian process model to fit time-series data from

microarray and establish a likelihood ratio test to rank differentially-expressed genes. Linear regression model is the easiest and the most widely used regression model, which is normally the foundation of other types of models. We begin with linear regression model by giving the definition

$$f(x) = x^T w \quad y = f(x) + \varepsilon$$

Where x is the input vector (for our time series data, x is the vector of time), w is a vector of parameters of the model, f is the function value and y is the observed target value (gene expression value in our case). By assuming the noise follows Gaussian distribution with zero mean and variance σ_n^2 and parameters follows Gaussian distribution with zero mean and variance σ_w^2 , we have

$$\varepsilon \sim N(0, \sigma_n^2)$$

$$w \sim N(0, \sigma_w^2)$$

The marginal likelihood function can be derived. Since it is jointly Gaussian, we have

$$p(y|x) \sim N(0, K_y)$$

where K_y is the variance (covariance) matrix. Then we have:

$$p(y|x) = \frac{1}{(2\pi)^{\frac{n}{2}} |K_y|^{\frac{1}{2}}} \times \exp\left(-\frac{1}{2}(y-m)^T K_f^{-1}(y-m)\right),$$

$$\text{or } y|x \sim GP(y; m(x), K_f(x_i, x_j))$$

where the mean function and covariance function are

$$m(x) = \langle f(x) \rangle,$$

$$K_f(x_i, x_j) = \langle (f(x_i) - m(x_i))(f(x_j) - m(x_j)) \rangle$$

This assumption can be derived when the model is a linear model. The Bayesian linear model assumes that the noise and parameters follow Gaussian distribution and the parameters are seen as the prior. Then based on Bayes' rule, marginal likelihood function can be derived.

Next step, the model chooses an appropriate covariance matrix K_y . No matter what covariance matrix is used, it must meet two requirements - Kolmogorov consistency and exchangeability. Kolmogorov consistency is the technical restriction on covariance function which means it must be positive semi-definite, that is, $y^T K y \geq 0$. It also needs to satisfy Stationary Gaussian Process, that is, $F_X(x_{t_1+\tau}, \dots, x_{t_k+\tau}) = F_X(x_{t_1}, \dots, x_{t_k})$. The function F_X remains the same when the input vector has a shift.

The Lawrence method chooses to use the most commonly-used kernel function, Squared-exponential kernel (SE), to calculate the variance/covariance matrix K_y . Log-marginal likelihood $\ln p(y|x, \theta)$ can be derived as below.

$$K_y(x_i, x_j) = \sigma_f^2 \exp\left(-\frac{(x_i - x_j)^2}{2l^2}\right) + \sigma_n^2 \delta_{ij}$$

Where σ_f^2 is the signal variance, l^2 is the characteristic length-scale and δ_{ij} is the Kronecker delta function. These three parameters are also called hyperparameters. Based on all the derivations above, we can obtain a log-marginal likelihood.

$$\ln p(y|x, \theta) = -\frac{1}{2} y^T K_y^{-1} y - \frac{1}{2} \ln |K_y| - \frac{n}{2} \ln 2\pi$$

The analysis of one-sample data which only involves one sequence uses a simpler approach rather than the full Bayes factor to rank the differentially expressed genes.

$$\ln\left(\frac{p(y|x, \theta_2)}{p(y|x, \theta_1)}\right)$$

In the above log-ratio of marginal likelihood function, θ_1 and θ_2 ($\theta = (l^2, \sigma_f^2, \sigma_n^2)^T$) are two hyperparameters based on two different hypotheses H_1 and H_2 . On hypothesis H_1 , Lawrence method assumes that there is no underlying signal of the gene profile, which means the gene profile is purely noise. Based on this assumption, hyperparameters can be determined with $\theta_1 = (\infty, 0, \text{var}(y))^T$.

$l^2 \rightarrow \infty$ indicates that the gene expression values of any two time points are irrelevant. Since all the gene expression are treated as noise, signal variance $\sigma_f^2 = 0$ and noise variance σ_n^2 can be decided by the gene profile, that is $\sigma_f^2 = \text{var}(y)$. While on hypothesis H_2 , Lawrence assumes that there is an underlying signal in the gene profile with no noise. Hyperparameter θ_2 can be initialized with $\theta_2 = (l^2, \text{var}(y), 0)^T$, in which l^2 can be fixed with any reasonable value to represent any two relevant time point distance, σ_f^2 is the variance of one gene expression profile and $\sigma_n^2 = 0$ since there is no noise.

Alternatively, in Stegle's method, the log-marginal likelihood is $\ln p(y_c, y_t | x, \theta)$, where y_c is the gene expression value of the control sequence and y_t is the gene expression value of treatment sequence. The log-marginal likelihood thus can be divided into two sequences:

$$\ln p(y_c, y_t | x, \theta) = \ln p(y_c | x, \theta_c) + \ln p(y_t | x, \theta_t)$$

where θ_c and θ_t represent the hyperparameters for the two profiles of the input data, respectively. And c and t are short for “control” and “treatment”, respectively.

Stegle's method deals with two hypotheses: null hypothesis and alternative hypothesis, namely H_0 and H_1 . H_0 means no difference between control and treatment gene expression value; H_1 means the profile differentially expressed on treatment group.

When the model is under the hypothesis H_0 , $\theta_c = \theta_t$. If the model is under H_1 , there is no requirement about theta; they can be the same or different. Therefore, the likelihood ratio reflects how likely the profile could be differentially expressed.

Mapping DE or Active Genes to KEGG Pathways

The combined DE and active genes were annotated using Blast2GO to remove genes with an E -value $> 10^{-3}$ (Götz et al., 2008). The Blast2GO-filtered genes were then mapped to KEGG pathways using RefNetBuilder previously developed by our lab (Li et al., 2011). The mapped KEGG pathways were ranked by the enrichment rate as determined by the percentage of the carbaryl- or RDX-affected GOIs in the total number of KEGG genes for each specific mapped pathway. Figure 11 shows GABAergic synapse pathways, one of the top 50 KEGG pathways enriched with carbaryl- or RDX-affected genes. Sixteen KOIDs were mapped to 15 nodes on the pathway. Multiple earthworm genes might be mapped to the same KEGG node. For instance, four or five earthworm genes were mapped to the KEGG gene "Gi/o" in the GRNs in Figure 16.

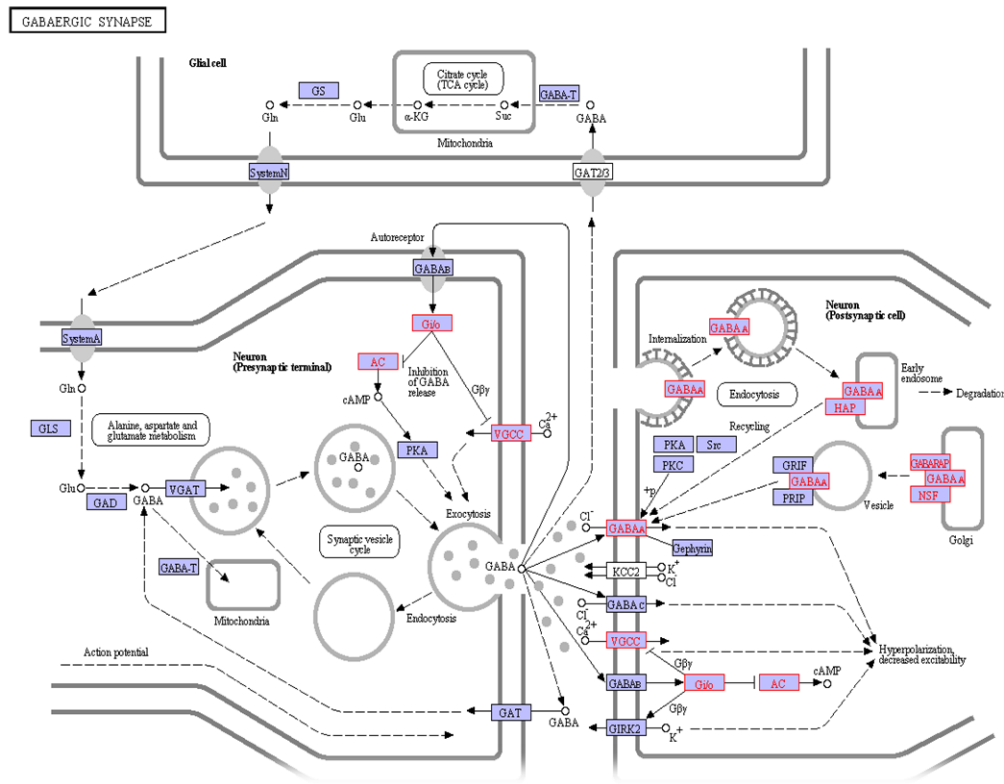


Figure 11. Mapped genes on GABAergic synapse pathway on KEGG.

Bayesian Learning and Optimization Model

The temporal expression data of genes mapped to 10 selected KEGG pathways were collated and used to reconstruct GRNs using the Bayesian Learning and Optimization model developed by the Computational Biology and Bioinformatics Laboratory (CBBL) at the University of Southern Mississippi for reconstruction and analysis of gene regulatory networks from time series gene expression data (Wu et al., 2011). Kalman Filter and Kalman Smoother were employed by BLOM to calculate gene interaction matrix.

Using the mapped DE or active genes from the above section as the input for BLOM, the reconstructed networks of all three treatments were then compared with the mapped KEGG pathway to infer differential networks (Yang, Maxwell et al., 2013).

Directed graphs were generated to show differential networks using open-source visualization tools Cytoscape and NodeXL (<http://nodexl.codeplex.com>). The perturbed gene interactions can be identified from the differential reconstructed gene networks.

Identification of Pathway Perturbations

Reversible neurotoxicity as measured by electrophysiological recording

No mortality occurred throughout the whole experiment. Conduction velocity of MGF suggests significant alteration during exposure to both RDX and carbaryl (Figure 12). At the end of the recovery phase, MGF's function was fully restored in all treatments. These results indicate the chosen concentration of both chemicals caused reversible neurotoxicity in earthworms.

Identification of significantly altered genes and pathways

The numbers of significantly altered genes at 27 time points for 2 treatments, as shown in Figure 13, suggest a reversible toxicity response of exposed earthworm to both chemicals. These results are in good agreement with the physiological response measured by MGF conduction velocity (Figure 12).

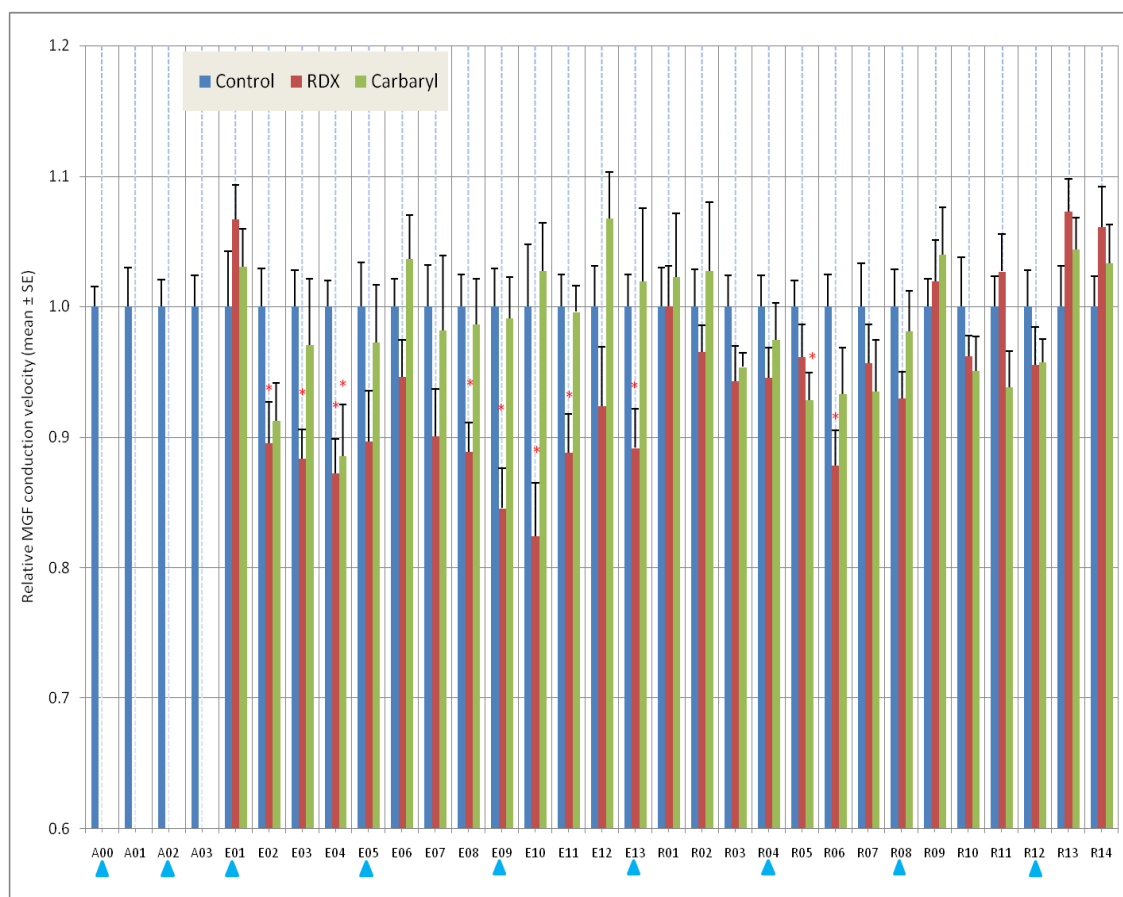


Figure 12. Effect of carbaryl or RDX exposure and recovery on earthworm MGF conduction velocity at 31 sampling time points over the course of 17 days. (Yang, Maxwell et al., 2013).

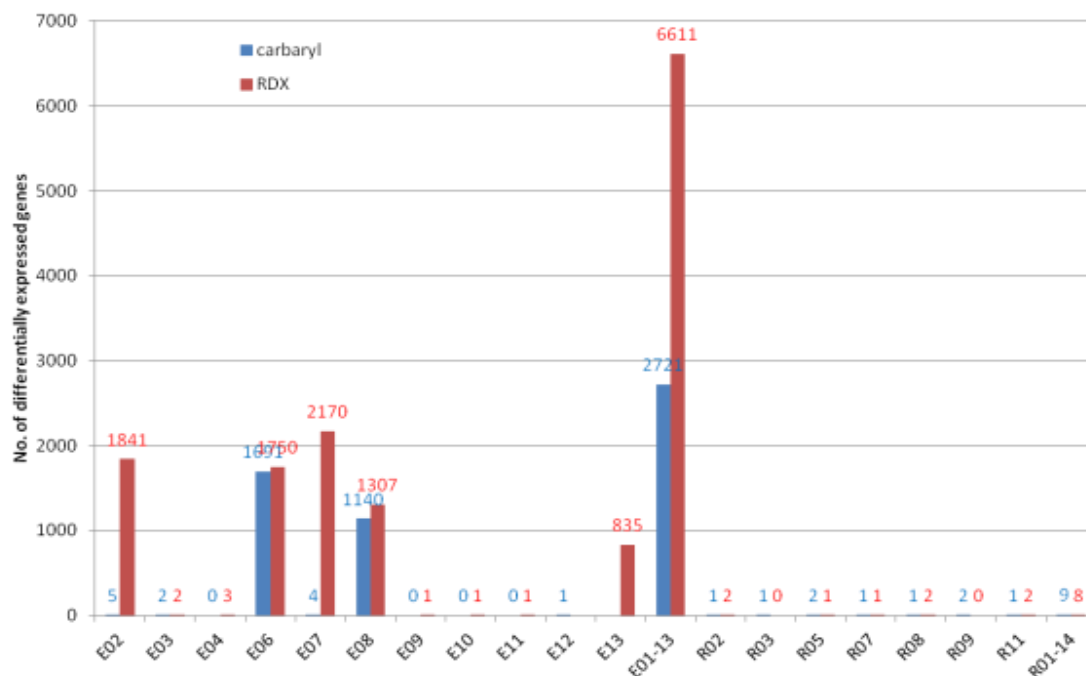


Figure 13. Number of DE genes identified at each individual sampling time point and the sum for the exposure (E01-13) and recovery (R01-14) phases.

None of the genes are differentially expressed at more than four time points in RDX-treated earthworms or more than two time points in carbaryl-treated earthworms (Figure 14). There were 1810 DE genes commonly affected by both chemical treatments. The intersections of top active genes ranked by Timecourse and GP2S are 737 and 4015 genes for carbaryl exposure and RDX exposure, respectively. There were no overlapping top-ranked active genes for carbaryl recovery and RDX recovery. The DE genes are combined with the overlap of top-ranked active genes to produce a list of genes of interest (GOIs) that amounted to 10715 unique genes (14099 with redundancy).

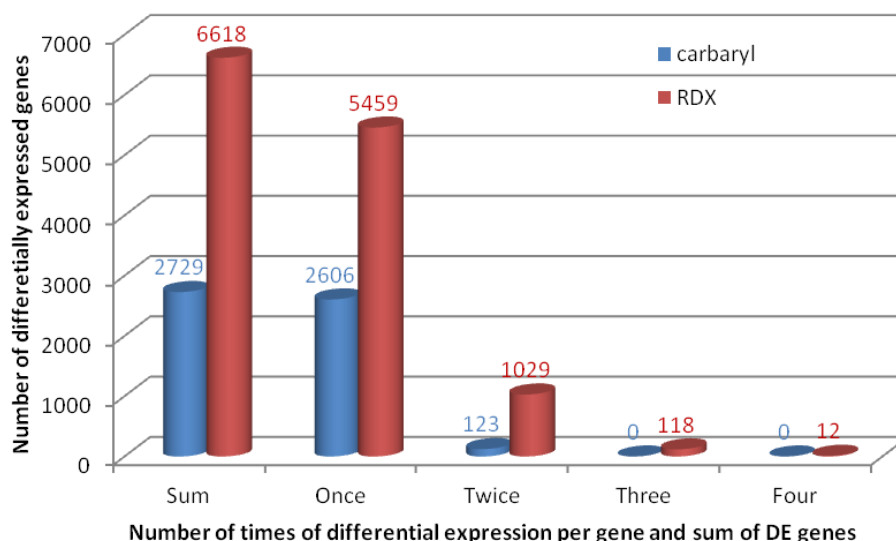


Figure 14. The frequency of differential expression of identified differentially expressed (DE) genes across 27 time points of exposure and recovery phases.

Blast2GO annotation of the 43803 probe-targeted transcripts reveals that 7423 (17%) have meaningful putative biological functions. Among these annotated transcripts, 2169 belong to GOIs (20% of 10715) with 734 genes being carbaryl-affected and 1941 RDX-affected. RefNetBuilder (Li et al., 2011) further mapped the 7423 annotated genes to 2529 KEGG genes in 224 pathways. Multiple earthworm genes might be mapped to the same KEGG node. For instance, four or five earthworm genes were mapped to the KEGG gene “Gi/o” shown in Figure 16. Among the mapped pathways, 169 were affected by carbaryl and 203 affected by RDX. The top 50 GOIs-enriched pathways (Figure 15) contain 8~27% and 19~47% KEGG genes mapped by carbaryl- and RDX-affected GOIs, respectively.

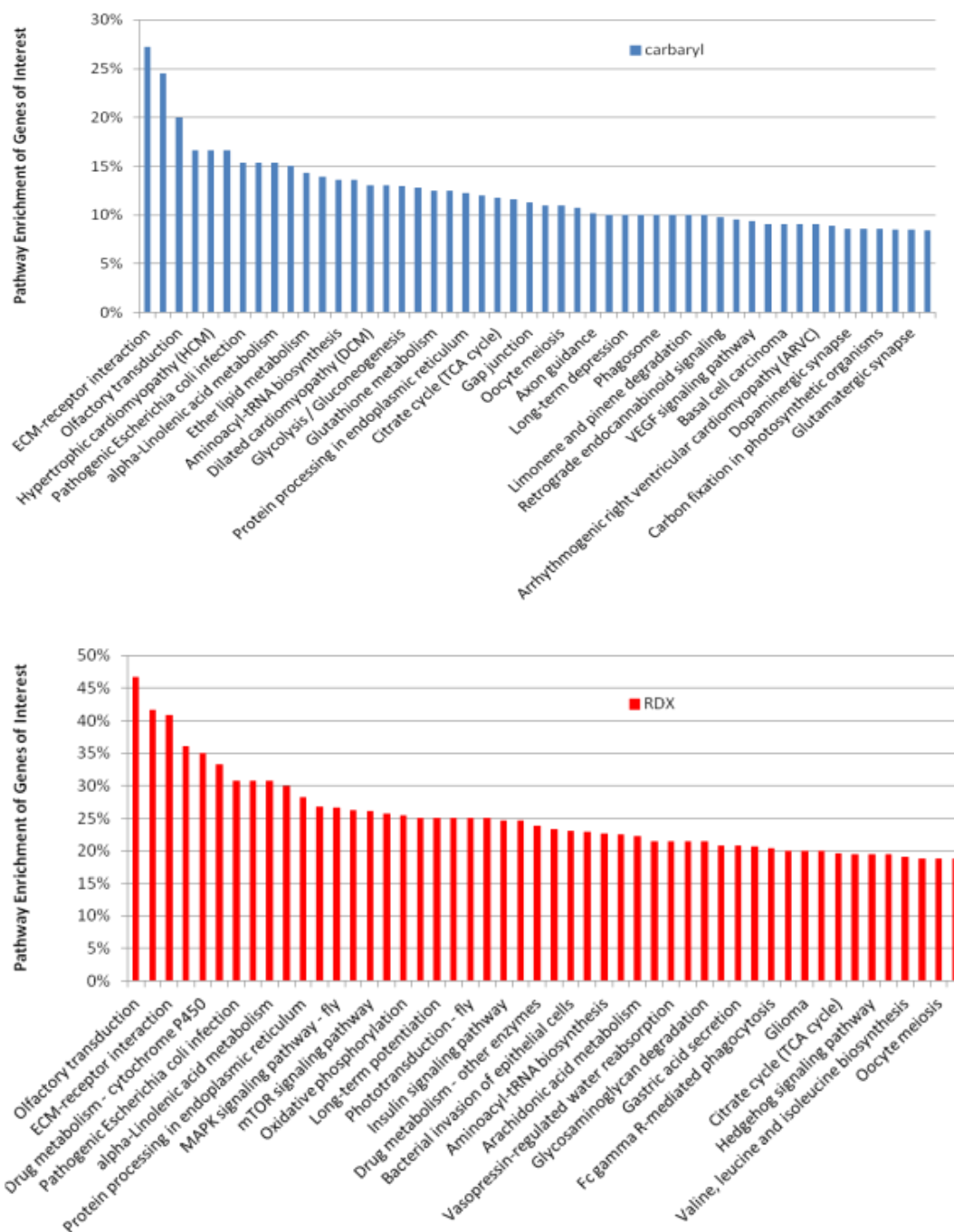


Figure 15. The top 50 KEGG pathways enriched with carbaryl- or RDX-affected genes.

A wide range of pathways have been affected by the two chemicals. They vary from olfactory transduction and Parkinson's disease to oocyte meiosis and axon guidance. Of the 10 nervous system-related pathways (koID: ko04720 to ko04730), long-term depression, dopaminergic synapse and glutaminergic synapse appear in the list of top 50 GOIs-enriched pathways affected by carbaryl, neurotrophin signaling is one of the top 50 pathways affected by RDX, and long-term potentiation is among the top 50 pathways affected by both RDX and carbaryl.

Differential reconstructed network to infer pathway perturbation

Expression data of earthworm genes mapped to the top GOI-enriched neurological pathways were used to reconstruct GRNs using BLOM model for three treatments and two phases. The number of edges in each of the six reconstructed networks was set to be the number of interactions between mapped genes existing in the canonical KEGG pathway. Using the DN approach detailed in Chapter III, differential edges, i.e., edges lost or gained from the control to the treated, were inferred from reconstructed networks and were used to construct differential networks.

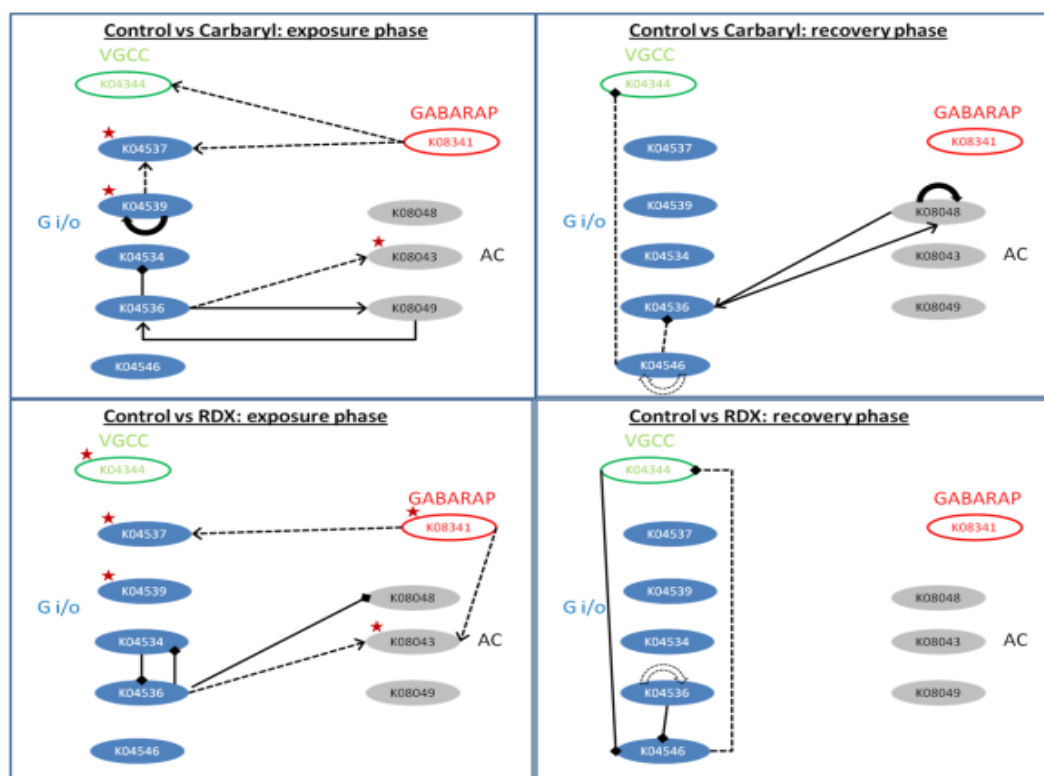


Figure 16. Differential networks (DNs) of the GABAergic synapse pathway consisting of differential edges inferred from pair-wise comparison of reconstructed GRNs between the control and RDX- or carbaryl-exposed earthworms. In the DNs, solid lines represent gained edges (edges absent in the control networks but present in carbaryl or RDX networks), whereas dashed lines represent lost edges (edges present in the control network but absent in carbaryl or RDX networks). Arrows indicate direction of gene regulation, while stars indicate identified GOIs. VGCC: voltage-dependent calcium channel, Gi/o: guanine nucleotide-binding protein, AC: adenylate cyclase, GABARAP: GABA_A receptor-associated protein.

A set of four differential networks derived for the GABAergic synapse pathway (Figure 16) show what gene interactions were altered by chemical treatments during the exposure phase and how some of the edges were later restored in the recovery phase (when the chemical was removed from the earthworms).

Pathway Perturbations

Although DE genes identified at a single time point are good candidates of biomarkers, they provide little interactive and dynamic information about multiple genes

involved in a pathway that are required to jointly carry out biological functions. A plausible solution is to collect temporal gene expression profiles, reconstruct GRNs from the time-series data, and infer differential gene interactions for select pathways by comparing reconstructed networks between the control and the treated.

Using a bioinformatics-guided reverse engineering approach, we have inferred differential GRNs that provide a close-up look of what interactions in an affected pathway might be perturbed. This study reinforces previous findings that cholinergic and GABAergic synapse pathways are the targets of carbaryl and RDX, respectively. RDX has been shown binding to the GABA_A receptor convulsant site, and blocking GABA_A receptor-mediated currents and causing seizures (Williams et al., 2011); carbaryl causes hyperstimulation of cholinergic receptors and an increase in excitatory neurotransmission (Jett, 2012). Several probes designed to target earthworm transcripts that putatively code for GABA receptors and cholinesterase were identified as GOIs (Yang, Li et al., 2013). However, cholinergic and GABAergic synapse pathways ranked in the 80s by enrichment analysis for both RDX and carbaryl, suggesting the existence of other targets.

Our results also indicate that perturbations to various pathways by sublethal concentrations of two neurotoxic chemicals were transient and recoverable. Many pathways other than the cholinergic and GABAergic synapse were altered during the exposure phase. Olfactory transduction and ECM-receptor interaction are the top two potential targets affected by RDX and carbaryl. They both warrant further in-depth investigations.

With the low meaningful annotation rates of the earthworm array (20%) and affected genes (17%), what we have discovered in the current study might have just been

the tip of an iceberg. A completely sequenced and annotated earthworm genome can empower the approach pursued in this study and will also aid future research.

CHAPTER V

DEVELOPMENT OF WEB-BLOM

Due to the complexity of mathematical models for GRN reconstruction, and the complexity of parameter setting and running commands involved in this process, it is time consuming and impractical for biologists to install the software packages on their local machines to use the models. Therefore, user-friendly web-based applications that can be accessed remotely by users facilitate GRN analysis. Web-BLOM is such a web-based software tool developed in this dissertation for reconstruction and analysis of gene regulatory networks from time-series gene expression data.

Workflow

The web site consists of several modular components: a database, GRN reconstruction model and a user interface. Figure 17 shows the workflow of this web site. The database is used to store user information, the data files uploaded by the users, and the results from the reconstruction model. The Bayesian Learning and Optimization Model (BLOM) is a model developed by the Computational Biology and Bioinformatics Laboratory (CBBL) in School of Computing for gene regulatory network reconstruction (Li, 2009; Wu et al., 2011). The BLOM, originally implemented in MATLAB, was adopted by the current version of Web-BLOM to provide online reconstruction of large-scale gene regulatory networks. Compared to other network reconstruction models, BLOM is much more computationally efficient.

Web-BLOM is designed in a three-tier architecture model (Figure 18) and implemented in MATLAB and Java. MATLAB Java Builder toolbox can package BLOM code into Java archive classes. Since MATLAB functions are wrapped into Java

classes and can be used within Java servlets, Web-BLOM is platform independent and can run on any standard computer. From the user interface, users can upload their time-series gene expression data to the server and manage all datasets. If the uploaded files pass an examination of the accepted size and file types, the users can search gene names in the uploaded file and then select a subset of genes. Then, the user interface remotely activates the BLOM that runs on a dedicated server. After Web-BLOM completes the submitted tasks, it generates a matrix of confidence values that can be used for inferring the interactions among selected genes.

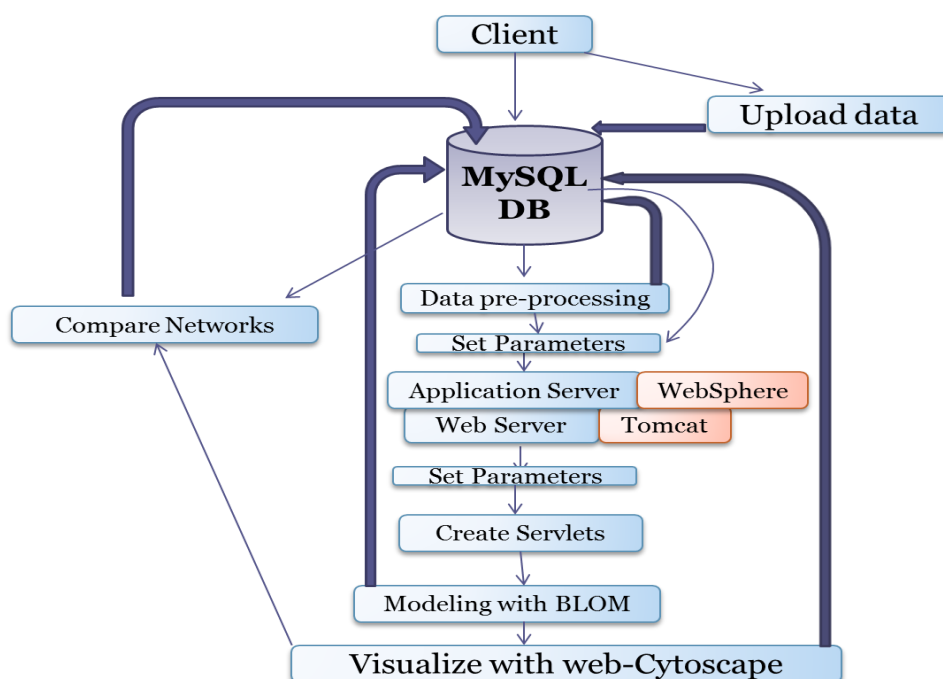


Figure 17. Workflow of Web-BLOM.

If users are interested in prioritizing genes for functional screening, they can use Web-BLOM to return a list of interacting genes ranked by confidence values. Web-BLOM integrates feature selection, network reconstruction, and result analysis in an

online network environment and it provides a new efficient and convenient software tool for reconstruction and analysis of gene regulatory networks.

Advantages of BLOM for a Web-Based Application

The limitations of many of the existing computational models to infer gene regulatory network lie in that they suffer from high computational complexity. Therefore, a lot of web-based applications based on those models, such as miniTUBA (Xiang et al., 2007), which is based on Dynamic Bayesian Networks, either return the results to users by email after computation, or only handle a small network. Compared with other GRN reconstruction models such as DBN, BLOM is much more computationally efficient (Li, 2009). Additionally, the output of BLOM provides more information about gene-gene interactions including the types of regulation (inhibition or activation), regulation directions (e.g. gene A regulates gene B or gene B regulates gene A) and strength of inferred interactions, which can be used as a parameter for ranking and belief management. In addition, BLOM can be used to reconstruct large-scale networks (Wu et al., 2011). My tests on different platforms showed that Web-BLOM can return the result for a 200-gene data set to different browsers in less than one minute (see details in the “Performance Test” section of Chapter V).

System Architecture

Traditionally, web-based applications have adopted a three-tier architecture where three independent layers or tiers, i.e. presentation, logic and data tiers, are configured (Kohl, Lotspiech, & Kaplan, 1997). The graphical user interface (GUI) and any other component in which a user interacts with the application are handled by the presentation tier, while the data tier manages the internal and external storage of application-related

data and provides access to it. The logic tier acts as a connector between the other two layers, handling their communication and performing any logical processing and analysis of data using various computational resources. The three-tier structure is dependent on the connectivity existing between all tiers to pass parameters. Web-BLOM adopts this traditional three-tier architecture.

BLOM was originally implemented in MATLAB, which provides a Java Builder toolbox to package MATLAB functions into a Jar file. The MATLAB functions wrapped in Java classes can be called by Java applications. Java servlet is chosen as the middle tier for Web-BLOM (Figure 18). Compared with an older technology, CGI (Common Gateway Interface) scripts, Java servlets are much more secure. CGI scripts can potentially impose some security issues (Kou & Springsteel, 1997). For example, Simultaneous CGI requests cause the script to be copied and loaded into memory as many times as there are requests. However, with Java servlets, there is the same amount of threads as requests, but there will only be one copy of the servlet class created in memory that persists between requests. Only a single instance answers all requests concurrently. This reduces memory usage and makes the management of persistent data easy (Pursnani, 2001).

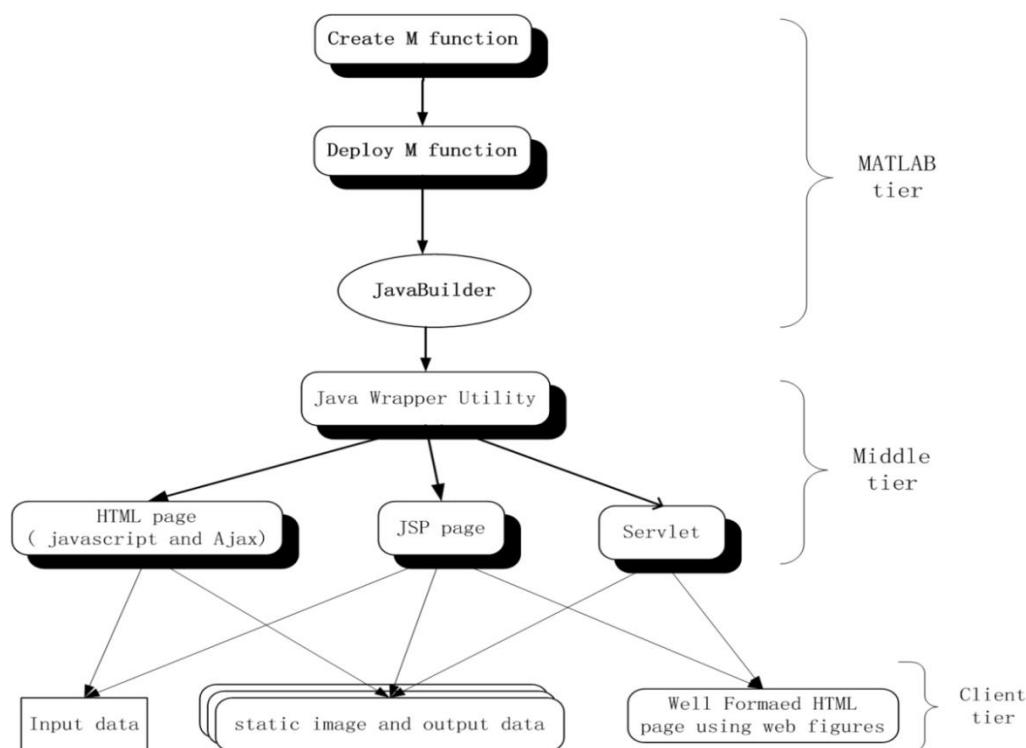


Figure 18. Three-tier architecture of Web-BLOM.

Aside from security issues, traditional CGI scripts written in Java, Perl or R have a number of disadvantages when it comes to performance. When an HTTP request is made, a new process is created for each call of the CGI script. This overhead of process creation can be very system-intensive, especially when the script does relatively fast operations. Thus, process creation will take more time than CGI script execution. Java servlets solve this, as a servlet is not a separate process. Each request to be handled by a servlet is handled by a separate Java thread within the Web server process, omitting separate process forking by the HTTP daemon.

A JavaScript library Cytoscape Web (Lopes et al., 2010) is used to visualize the gene interaction matrix into a GRN. Cytoscape Web was implemented in Flex and ActionScript. Figure 19 shows its architecture which has the advantage of using the Flash platform to implement complex and interactive vector images that behave consistently

across major browsers, but without requiring the web site to be entirely built with this technology.

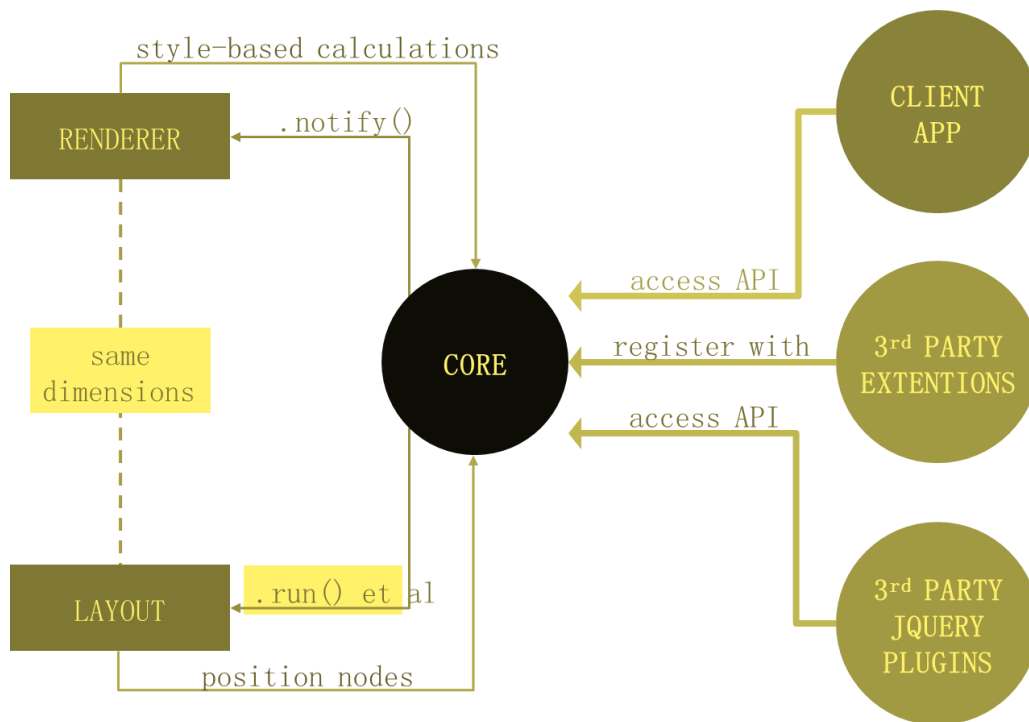


Figure 19. Architecture of the JavaScript library, Cytoscape Web, which was used by Web-BLOM (Lopes, 2010).

Implementation

Implementation Environment

The web-BLOM site was built on two virtual machines created by VMware Hypervisor ESXi 4.0 on a Dell PowerEdge M905 physical server in School of Computing of the University of Southern Mississippi. The database of Web-BLOM is installed on the virtual server - bluefin, for which 2G of memory is allocated. MySQL 5.0.77 is installed on bluefin as the database management system. The database is used to store user data, uploaded data and curated links to other network analysis tools. The user interface and servlets are stored on the other virtual server, “tc1”, for which 64G of memory is

allocated. Apache Tomcat Server 6.0 (<http://tomcat.apache.org>) is installed on “tc1” as the container for the servlets as well as JSP and HTML pages.

The Dell PowerEdge M905 physical server consists of 4 quad-core AMD processors, 72GB of RAM and 10Gbit Ethernet interconnects. CentOS release 5.7 is the operating system installed on the server. CentOS is an Enterprise-class Linux Distribution derived from sources freely provided to the public by an Upstream OS Provider (UOP).

Implementation of Servlets

MATLAB Java Builder (www.mathworks.com/products/javabuilder/) is an add-on toolbox to MATLAB that allows conversion of M-code into a JAR archive. We use MATLAB R2010b JA Builder to generate the jar file. It contains BLOM-related classes converted from BLOM’s MATLAB code, an MCR factory class generated by MATLAB Compiler Runtime and a manifest file, which is a metadata file that contains name-value pairs specifying the main class of the application. Building a JAR file from MATLAB .m files involves the following steps: (1) MATLAB-> File -> New -> deployment project; (2) rename the project to “blom.prj”; (3) set the directory you want to store the generated jar file; (4) select target as “Java Package”; (5) open deployment tool window (MATLAB-> desktop -> deployment tool) and click “add class” to create a new class “blom_class”; (6) build. After project building is finished (it takes a few minutes), a jar file is generated in the directory we set in the first step.

After the BLOM code was wrapped by MATLAB Javabuilder in the previous step, the servlet can now call the functions in the class files inside the jar archive. The following line of Java code shows how the servlet calls a MATLAB function.

The `outputArray` returned by this function call is an `MWArray` variable.

`MWArray` is a thin wrapper around a MATLAB array and can be converted into a 2-D array in Java for output. During the initialization stage of the Servlet life cycle, the web container initializes the servlet instance by calling the `init()` method as shown in Figure 20. The container passes an object implementing the `ServletConfig` interface via the `init()` method. This configuration object allows the servlet to access name-value initialization parameters from the web application.

After initialization, the servlet can service client requests. Each request is serviced in its own separate thread. The web container tomcat calls the `service()` method of the servlet for every request. The `service()` method determines the kind of request being made and dispatches it to an appropriate method to handle the request. The developer of the servlet must provide an implementation for these methods. If a request for a method that is not implemented by the servlet is made, the method of the parent class is called, typically resulting in an error being returned to the requester. Finally, the web container calls the `destroy()` method that takes the servlet out of service (Figure 20). The `destroy()` method, like `init()`, is called only once in the lifecycle of a servlet.

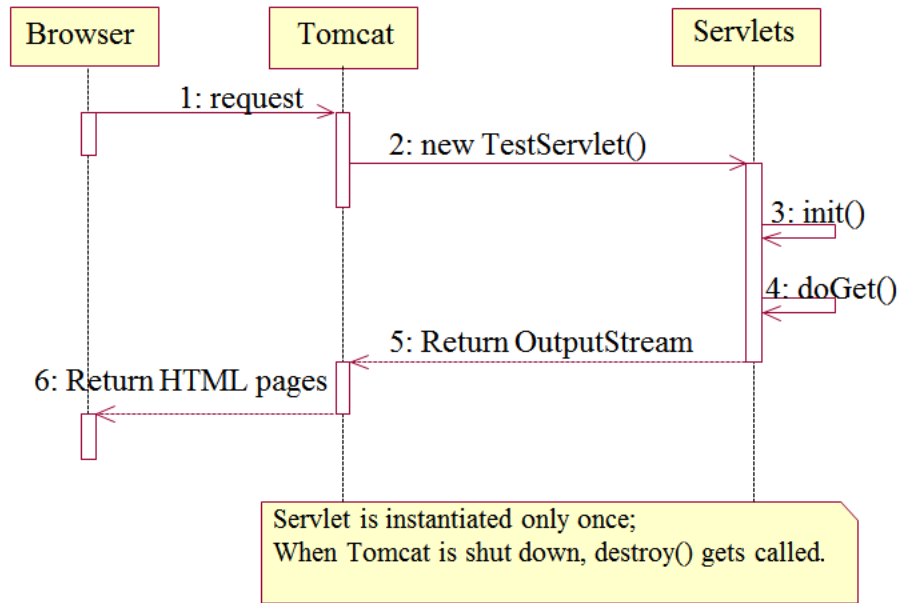


Figure 20. Workflow of a servlet in Apache Tomcat.

POI (Poor Obfuscation Implementation) is an apache API which provides methods for the servlet to read and process Microsoft Excel files. After downloading POI package from the apache website (<http://poi.apache.org/>), the 5 jar files in POI's home directory need to be copied to both Tomcat's lib directory and java JRE's extended library directory (.../jre/lib/ext) for compiling the servlet code later on. MATLAB's javabuilder.jar also needs to be copied to these two directories. Alternatively, we can add the location of these jar files into the classpath when we compile the servlet. For example, we can compile the servlet with the following command:

```
javac -classpath "./WEB-INF/lib/blom.jar" BlomVarArgServlet.java
```

To successfully compile the servlet on a Linux server, we need to set the path and classpath by adding the following three variables into the ~/.bashrc file:

```
JAVA_HOME=/usr/local/jdk, PATH= $JAVA_HOME/bin:$PATH, CLASSPATH=
$JAVA_HOME/jre/lib/blom.jar. After compilation, three class files are generated and
```

they need to be moved to Tomcat's application classes directory (tomcat/webapps/blom2/*WEB-INF/classes/*) to be recognized as servlets by Tomcat.

A Case Study of Web-Based Rankprod

Rankprod is an R bioconductor package for detecting DE genes in meta-analysis (<http://www.bioconductor.org/packages/2.12/bioc/html/RankProd.html>). It is based on the statistical assumption that under the null hypothesis that the order of all items is random, the probability of finding a specific item among the top r of n items in a list is $RP=r/n$ (Hong et al., 2006). The smaller the RP value, the smaller the probability that the observed placement of the item at the top of the lists is due to chance.

Web-based Rankprod is developed using a Perl script to call the original R package of Rankprod stored on the same server as the Perl script. A microarray data set containing 55,515 probes (Walia et al., 2005) is used in web-based Rankprod to explore the transcriptome of the salt-tolerant and salt-sensitive rice genotypes in this study. The data set was downloaded from NCBI Gene Express Omnibus (Source: <http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE3053>).

The experiment that resulted in this data set was conducted by under control and salinity-stressed conditions during vegetative growth. Two rice genotypes, FL478, a recombinant inbred line derived from a population developed for salinity tolerance studies, and IR29, the sensitive parent of the population, were selected for this study. I applied Rankprod to this data set to compare their transcriptional profiles to identify DE genes. DE genes in salt-tolerant mutant type FL478 compared to its salt-sensitive wild type IR29 under salinity-stressed conditions. RiceCyc database (Source: <http://www.gamene.org/pathway/>) is a sub-database of BioCyc that stores manually-

curated rice metabolic pathways (Liang et al., 2008). RiceCyc can be downloaded from GRAMENE's ftp (<ftp://ftp.gramene.org/pub/gramene/>).

Using BioCyc's web application "Omics Viewer", 31 DE genes in wild type rice and 20 genes in mutant type rice were mapped to the cellular metabolism overview of RiceCyc. Meanwhile, the log value expression data are overlaid to the map. Pathway Tools calculates the fold change of the treatment to the control and displays a color gradient for the fold change, as shown in Figure 21 and Figure 22. The DE genes are mapped to 18 and 10 rice metabolic pathways (one gene can be present on multiple different pathways) in wild type rice and mutant type rice, respectively. In the IR29 genotype, DE genes are mapped to those pathways related to biosynthesis of plant hormones, cellular structural components and secondary metabolites. Each node in the map represents a metabolite which participates in the reaction and each edge represents the enzyme that catalyzes this reaction. Therefore, most of the DE genes were mapped onto the edges because the protein products of these genes act as enzymes of the mapped metabolic reactions. Some of the genes' protein products actually participate in the metabolic reactions themselves. Those genes that are located on the border of the map participate in the trans-membrane activities. The right side of the map shows standalone reactions that do not belong to any pathways. Pathways are defined as a chain of reactions (Schuster, Fell, & Dandekar, 2000).

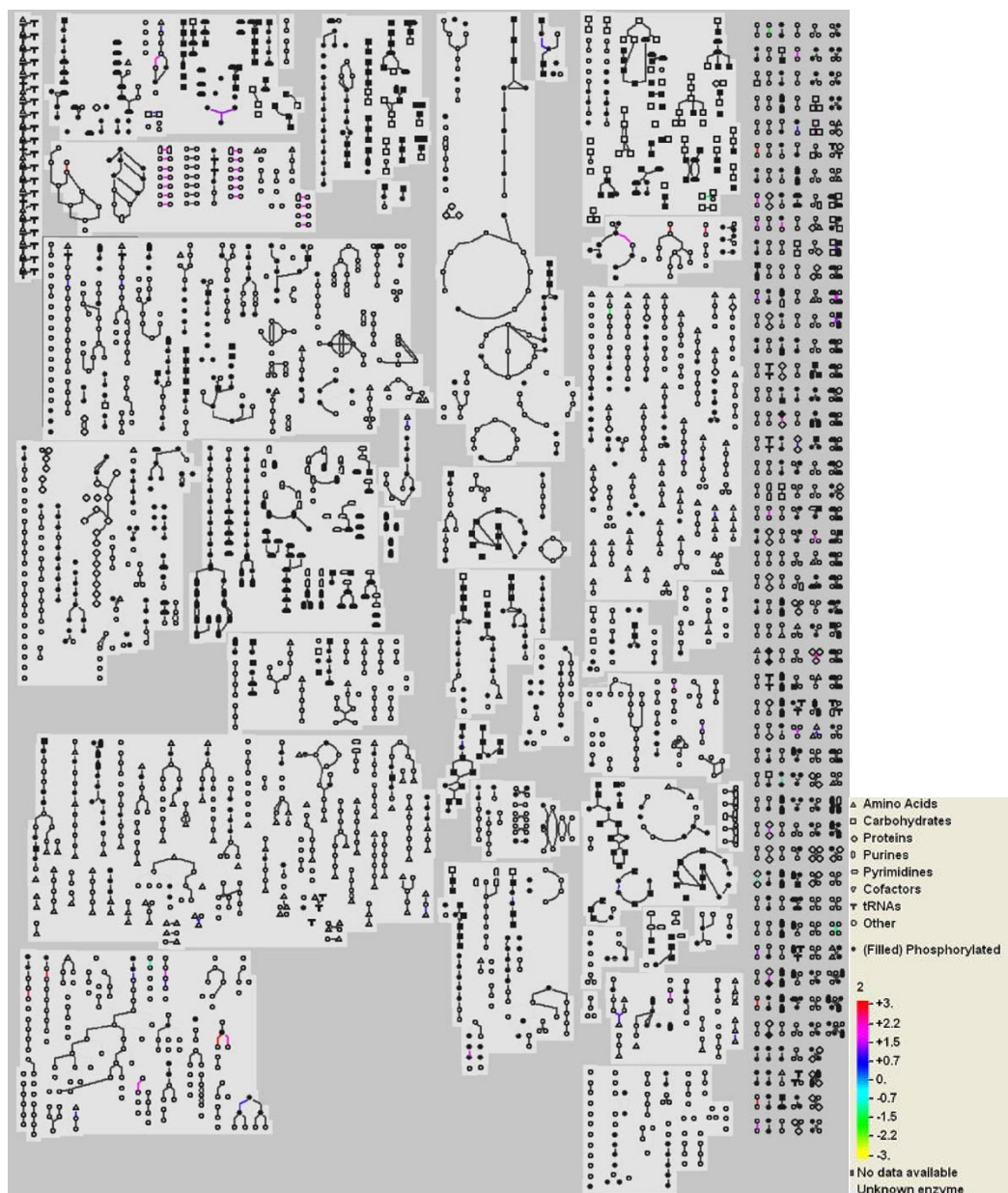


Figure 21. Differently-expressed genes of wild-type rice on global pathway diagram in Omics Viewer.

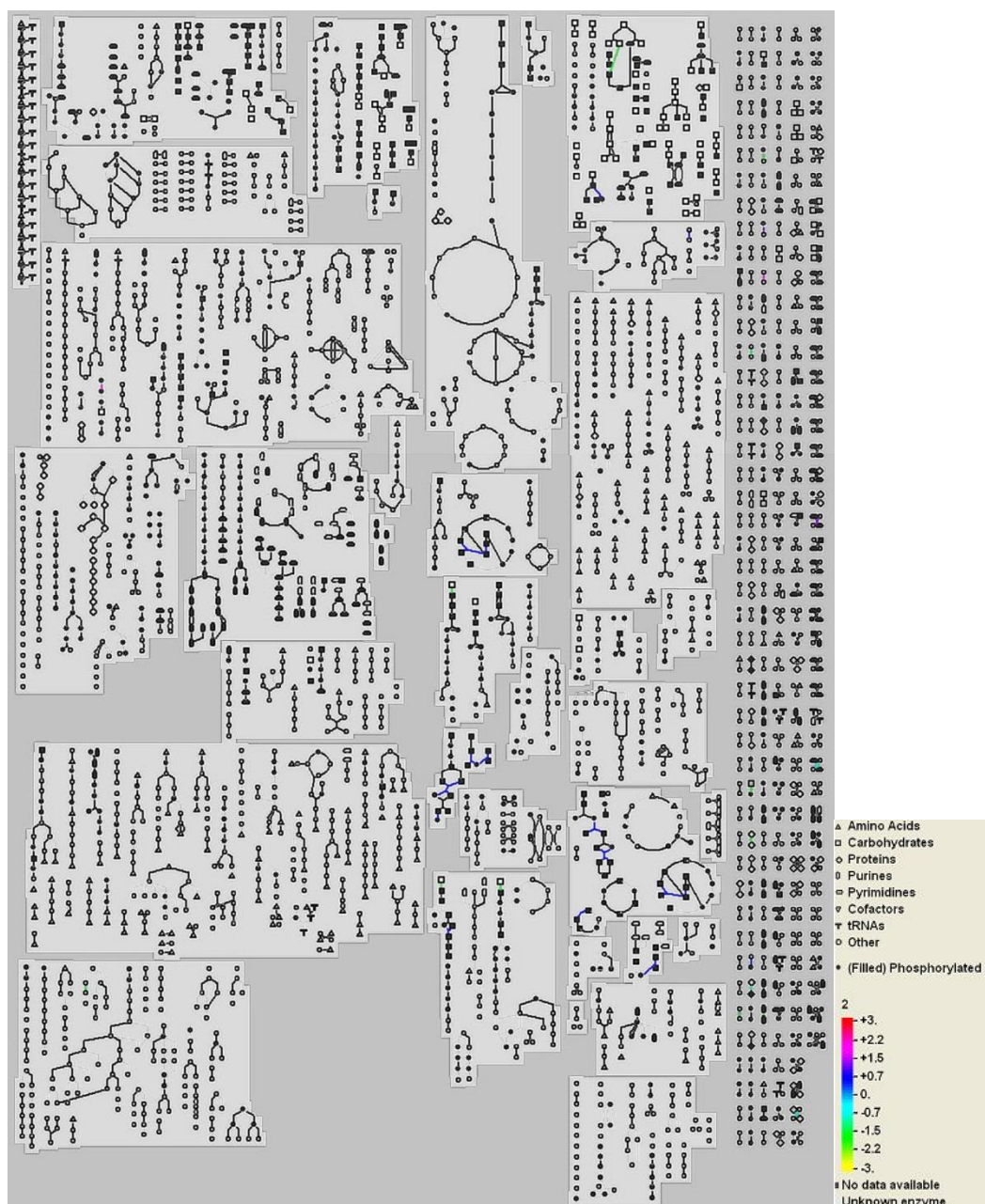


Figure 22. Differently-expressed genes of salt-enduring rice on global pathway diagram in Omics Viewer.

The global maps from Omics viewer (Figure 21 and Figure 22) showed that the DE genes identified by Rankprod are involved in energy recycling pathways in rice such as tri-carboxylic acid cycle (TCA cycle), pentose phosphate pathway, and ribose pathway in salinity-stressed FL478 genotype. Multiple peroxidase proteins are among the DE

genes in IR29, but not in FL478, which is consistent with the original discovery by the author of the data set (Walia et al., 2005).

A Case Study of Web-BLOM

User interface

Figure 23 shows the user interface of Web-BLOM. On this page, users can upload the data set to the server. The data set for uploading need to follow a specific format: each row is one gene and each column is a time point. Optionally users can then upload a list of genes of interest as an input file for web-BLOM to select a subset of data from the data file that the user uploaded under the condition that the first column of the uploaded data set contains the same identifiers as the gene list. This subset of data is then used for GRN reconstruction.

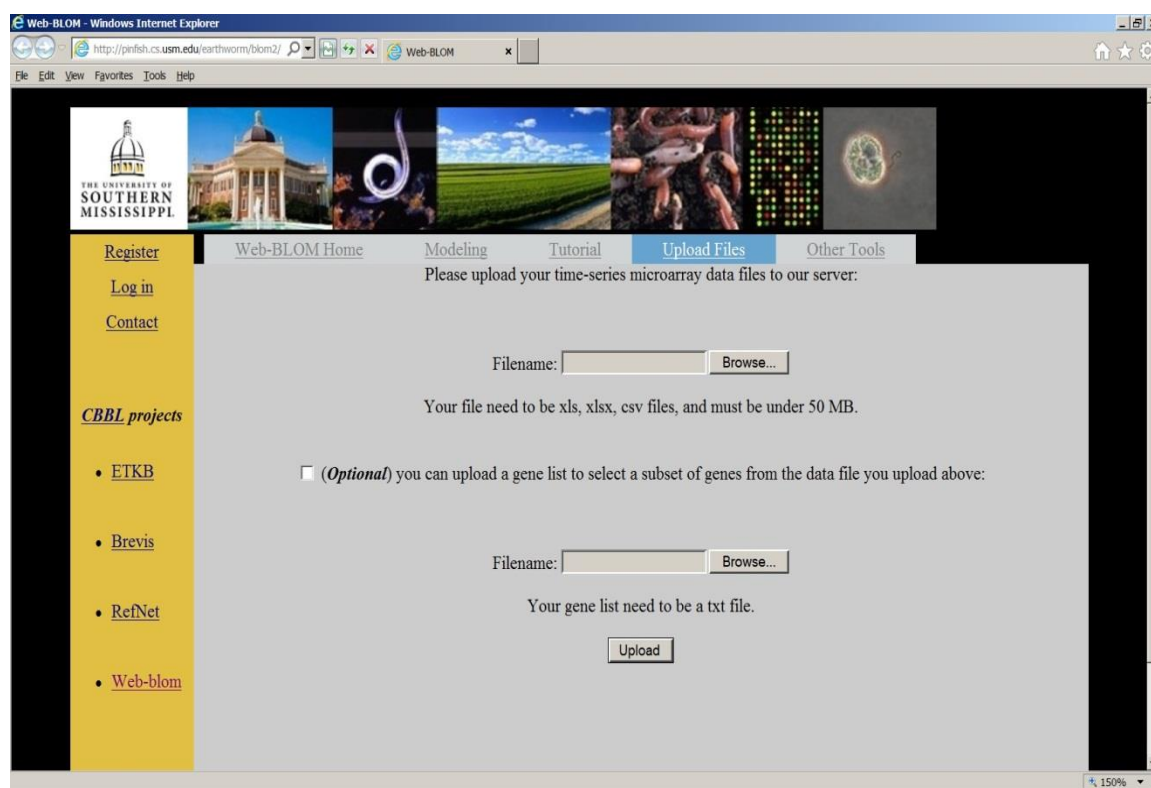



Figure 23. Web-BLOM page for uploading sorted gene expression data.

After uploading, four possible occasions might occur. (1) File uploaded successful and go to the parameter selection page. (2) If the file type is not xls or xlsx, the page will display the message: invalid file type. (3) If the file size exceeds 20MB, the page will display the message: the file exceeds the 20MB size limit. (4) If the name of the file is the same as an existing file in /data/ directory, then the embedded JavaScript on this page will append a time stamp at the end of the file name and the user will then be re-directed to the parameter selection page. Then users can select a subset from the full data set by determining the row numbers of the starting and ending genes as shown in Figure 24.



[Register](#)
[Log in](#)
[Contact](#)

[CBBL projects](#)
 • [ETKB](#)
 • [Brevis](#)
 • [RefNet](#)
 • [Web-bloom](#)

Modeling with the BLOM algorithm

Please Select a Data File you uploaded:

Or go back to the [uploading page](#).

You can select a subset of genes from the file you selected above by entering below the beginning and ending row and column numbers, or by [searching](#) gene names (gene names have to be located in the first column of your file).

Please enter the beginning and ending row numbers:

| | | | |
|----------------|----------------------|--------------|----------------------|
| Starting Row # | <input type="text"/> | Ending Row # | <input type="text"/> |
|----------------|----------------------|--------------|----------------------|

Please enter the starting and ending column letters (A-Z, AA-ZZ...):

| | | | |
|-------------------|----------------------|-----------------|----------------------|
| Starting Column # | <input type="text"/> | Ending Column # | <input type="text"/> |
|-------------------|----------------------|-----------------|----------------------|

Optional Parameters (you can skip all of these and jump directly to [submit](#).)

Numerical Parameters

Maximum number of edges in the output
 Note: This number can range from 0 to num_gene^2.

diagQ (enter 0 or 1)
 diagR (enter 0 or 1)

Matrix Parameters

Please input each number seperated by comma, and each row by semi-colon.

F0 (see below)
 H0 (see below)
 Q0 (see below)
 R0 (see below)
 lower (see below)
 upper (see below)

Note: It might take a few seconds to a minute to get the returned result depending on the size of the data. Do NOT refresh this page.

Figure 24. Web-BLOM page for users to select input parameters.

It might take a few minutes before a result matrix of confidence numbers are displayed, depending on the size of the input data (see section 5.6.4 for performance testing results). The user might click the “submit” button again or even refresh the page while the servlet is still performing the algorithm. Either refreshing or multiple clicks of “submit” might lead to failure to displaying the result. Therefore, JavaScript code is embedded in this page to forbid refreshing.

Output

After the servlet completes the submitted tasks, it generates a matrix of confidence values (range from -1 to 1) that indicates the interactions among selected genes. The result page also returns a matrix of the input data from the genes and time points the user selected in the query page, for the user to verify the correctness of their data selection. A sample output page is shown in Figure 25. The earthworm microarray data set used in Figure 25 is the same data set used in Chapter IV and a random set of 5 genes is selected from this 43803-gene data set for demonstration. The positive values indicate enhancing regulations while negative values indicate inhibitory regulations. If users are interested in prioritizing genes for functional screening, they can select pairs of genes with highest absolute confidence values because the possibility of regulation relationships is higher among these pairs of genes. See Chapter IV for cut-off value selection techniques. Figure 26 shows the visualized network using the JavaScript library Cytoscape Web (Lopes et al., 2010).

Input Data

5 genes & 5 time points

| ProbeName | E01 Carbaryl | E02 Carbaryl | E03 Carbaryl | E04 Carbaryl | E05 Carbaryl |
|------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| TA1-108326 | 0.374165 | 0.85349 | 0.683062 | 0.252377 | 0.559765 |
| TA1-059409 | 3.093235 | 2.98573 | 3.076397 | 3.105754 | 3.180328 |
| TA2-099409 | -0.15679 | -0.79425 | -0.24438 | 0.052943 | -1.01988 |
| TA2-018956 | -0.15802 | -0.58977 | 0.138595 | -0.54983 | 0.070498 |
| TA1-163303 | 1.123192 | 1.065739 | 1.425972 | 1.049741 | 0.895641 |

Output Connection Matrix

| gene IDs | TA1-108326 | TA1-059409 | TA2-099409 | TA2-018956 | TA1-163303 |
|------------|------------|------------|------------|------------|------------|
| TA1-108326 | 0.363 | 0.210 | -0.065 | 0.008 | -0.339 |
| TA1-059409 | 0.020 | 0.223 | -0.103 | 0.074 | -0.092 |
| TA2-099409 | 0.088 | -0.201 | 0.163 | 0.413 | 0.693 |
| TA2-018956 | 0.047 | -0.198 | 0.511 | -0.009 | -0.328 |
| TA1-163303 | -0.008 | 0.121 | 0.001 | 0.005 | 0.402 |

Figure 25. Web-BLOM result page that returns both the input matrix from the genes and time points that the user selected, and an output matrix of confidence values.

Result from Web-BLOM

| | TA2-099409 | TA2-018956 | TA1-163303 | TA |
|--------|----------------------|----------------------|---------------------|-----|
| 002376 | -0.11457605543458871 | -0.33511203563488867 | 0.9093341328227385 | -0. |
| 000182 | 0.12754157933308163 | -0.5192902306677771 | -0.4238027168234182 | -0. |
| 913697 | -0.06307629656103031 | 0.8300138917769886 | -0.8054972058043899 | 0.1 |
| 635518 | 0.977058729400275 | 0.5026036936714815 | 0.190531119227759 | 0.5 |
| 24282 | 0.2770754676478946 | 0.7655523767029988 | 0.03071081376251117 | -0. |
| 73214 | 0.967630273926503 | -0.4208437183248479 | -0.939663831698597 | 0.6 |
| 68686 | 0.28206206731693806 | 0.019566480785912166 | 0.83866714742535 | 0.9 |

Please input a cut-off value:

[Export the output matrix to a txt file](#)

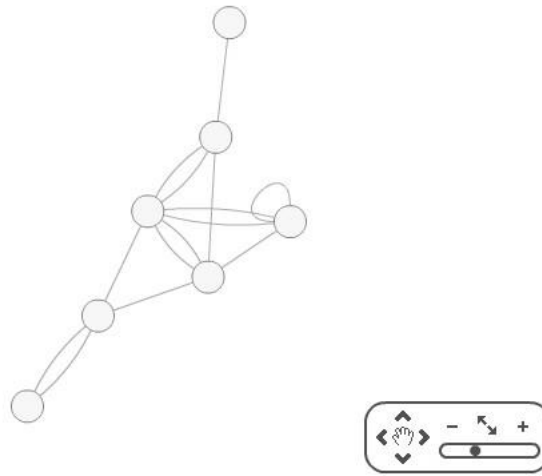


Figure 26. Visualization using a JavaScript library, Cytoscape Web.

Performance testing

Web-BLOM works best with small to medium-sized networks, generally with up to a few hundred genes. Larger networks can be calculated, but the user interaction can become sluggish around 300 genes (nodes or edges). A second factor is the size of the original uploaded data file. When the file size is as large as 100MB, then it might take a few minutes to perform the algorithm. Two test cases of a few hundred genes with 13 and 31 time points from a 20MB excel file were performed and the runtime that Web-BLOM took to generate the result matrices on different browsers is shown in Table 6.

Table 6

Time in seconds for web-BLOM to return the results for two different large-sized networks from a 20MB excel file on three different browsers, tested on a ThinkPad laptop with 2 GHz dual core CPU and 2 GB RAM.

| Browsers | # of time points | 100 | 200 | 300 |
|---------------------|------------------|-----|------|-------|
| Firefox 4.0 | 13 | 2.5 | 29.5 | 80.0 |
| | 31 | 2.5 | 38.0 | 302.0 |
| Chrome 10 | 13 | 2.5 | 29.0 | 79.0 |
| | 31 | 2.5 | 37.5 | 303.5 |
| Internet Explorer 8 | 13 | 4.5 | 30.5 | 81.0 |
| | 31 | 4.5 | 42.0 | 305.0 |

CHAPTER VI

CONCLUSIONS AND FUTURE DIRECTIONS

Conclusions

The proof-of concept study on LCA in this dissertation proposed a DNs approach to analyzing time-series gene expression datasets and connected pathway perturbation with toxicity threshold setting. The DNs approach proposed in this dissertation differs significantly from the existing DE genes-based approaches such as the Gene Set Enrichment Analysis (GSEA) method (Subramanian et al., 2005) and the benchmark dose method in at least three aspects: (1) this approach is based on reverse engineering techniques including BLOM, PBN and DBN; (2) significantly altered pathways are identified through analysis of DNs instead of enrichment of DE genes mapped to canonical pathways; and (3) this approach is particularly suitable for analyzing time-series gene expression datasets whereas existing approaches like GSEA are suitable for static datasets that are often collected at a single time point. Gene expression data at one single time point have limited power in both deciphering MOAs and quantitative risk assessment because the snapshot of gene expression profiling misses the dynamic and interactive nature of cellular gene expression. As the costs of acquiring genome-wide gene expression technologies steadily decrease, it has become more feasible and affordable to perform time-series gene expression studies. In order to take advantages of technological advancements in high throughput microarray, DNA sequencing and LCA, novel experimental and computational approaches are needed to transform conventional toxicology to predictive toxicity in order to meet the requirements of more rapidly assessing toxicity of chemicals and other materials to humans and animals in the 21st

century risk assessment (Bhattacharya et al., 2011; Collins et al., 2008; Cote et al., 2012; Currie, 2012; Krewski et al., 2010; Tannenbaum, 2012).

A distinction has to be made between edges in the graphical representation of a literature curated biological pathway (e.g., a KEGG pathway) and those derived in silico from time-course data. The former are experimentally validated interactions, whereas the latter represent potential gene-gene interactions. It was not our intention to estimate the accuracy of our BLOM-inferred edges by comparing them with those gene-gene interactions curated in KEGG pathways or EcoCyc databases, but rather to use the inferred edges to provide an estimate of the overall degree of alteration in a gene's interaction/connectivity with other genes.

It has also to be noted that in the current study there exist the following four limitations. (1) No chemistry work was carried out to confirm the concentration and/or biotransformation of NAs throughout the 3-hr exposure. Many xenobiotic toxicants such as polycyclic aromatic hydrocarbons, aryl and heterocyclic amines require metabolic activation by cytochrome P450s metabolism (Shimada et al., 2013). Chemical analysis, in parallel to bioassays, can provide useful information about the chemical(s) of concern for toxicity threshold derivation. (2) The absence of treatment replication in addition to the low treatment number made pathway perturbation degrees (Table 2) practically inadequate to statistically determine a point of departure or toxicity threshold for each perturbed pathway. Apparently, this limitation can only be ameliorated by collecting time-course datasets with more treatments and treatment replications. (3) No apical endpoints such as cytotoxicity, physiology or biochemistry assays were measured, making the derived lowest observable pathway perturbation concentrations one step

shorter from being correlated to toxicity thresholds derived from apical endpoints and hence applicable to chemical risk assessment (Thomas et al., 2012; Thomas et al., 2013).

(4) Despite the aforesaid advantages, the *E. coli* LCA system is not free from limitations. For instance, less than 50% of known transcriptional genes have a promoter that can be fused with a GFP (Zaslaver et al., 2006), leading to an incomplete genome coverage. Alternative high-throughput technologies such as DNA microarray and next-generation sequencing can be used to generate genome-wide time-series gene expression datasets.

Findings from this study suggest that our approach has a great potential in providing a novel and sensitive tool for threshold setting in chemical risk assessment. In future work, we plan to analyze more time-series datasets with a full spectrum of concentrations and sufficient replications per treatment, and eventually extrapolate our approach from prokaryotic systems to eukaryotes. The pathway alteration-derived thresholds will also be compared with those derived from apical toxicology, biochemistry, and physiology endpoints such as cell growth rate.

This DN's approach was later modified in Chapter IV to satisfy the need of studying earthworm pathways. Using a bioinformatics-guided reverse engineering approach, we have inferred from earthworm microarray data differential GRNs that provide a close-up look of what interactions in an affected pathway might be perturbed. This study reinforces previous findings that cholinergic and GABAergic synapse pathways are the targets of carbaryl and RDX, respectively. RDX has been shown binding to the GABA_A receptor convulsant site, and blocking GABA_A receptor-mediated currents and causing seizures (Williams et al., 2011); carbaryl causes hyperstimulation of cholinergic receptors and an increase in excitatory neurotransmission (Jett, 2012). Several

probes designed to target earthworm transcripts that putatively code for GABA receptors and cholinesterase were identified as GOIs. However, cholinergic and GABAergic synapse pathways ranked in the 80s by enrichment analysis for both RDX and carbaryl, suggesting the existence of other targets. The results also indicate that perturbations to various pathways by sub-lethal concentrations of two neurotoxic chemicals were transient and recoverable. Many pathways other than the cholinergic and GABAergic synapse were altered during the exposure phase. Olfactory transduction and ECM-receptor interaction are the top two potential targets affected by RDX and carbaryl. They both warrant further in-depth investigations. With the low meaningful annotation rates of the earthworm array (20%) and affected genes (17%), what we have discovered in the current study might have just been the tip of an iceberg. A completely sequenced and annotated earthworm genome can empower the approach pursued in this study and will also aid in future discovery journey.

The tests on different platforms showed that Web-BLOM can return the result for a 200-gene data set to different browsers in less than one minute. Web-BLOM is designed in a three-tier architecture model and implemented in MATLAB and Java. MATLAB Java Builder toolbox can package BLOM codes into Java archive classes. Since MATLAB functions are wrapped into Java classes and can be called by Java servlets, Web-BLOM is platform independent and can run on any standard computer. It generates a matrix of confidence values that can be used for inferring the interactions among selected genes. If users are interested in prioritizing genes for functional screening, they can use Web-BLOM to return a list of interacting genes ranked by confidence values. Web-BLOM integrates feature selection, network reconstruction, and result analysis in an

online network environment and it provides a new efficient and convenient software tool for reconstruction and analysis of gene regulatory networks.

Future Directions

The use of live cell arrays allows for reagentless, non-destructive real-time monitoring of the biological effects of chemicals. However, data pre-processing is very important in handling LCA data. When time-series OD₆₀₀ data is available from a Live Cell Array, it will further facilitate noise filtering. Recently, a software tool based on the discrete Kalman filter was developed by (Aichaoui et al., 2012) to provide standardized treatment to LCA data and generate reports on the quality of the data. For example, some certain types of bacteria such as *Bacillus subtilis* generate auto-fluorescence in the culture and this could become a background noise to the gene expression data. If there is no auto-fluorescence ($F_{\text{auto}} = 0$), the promoter activity is directly related to the time derivative of the fluorescence divided by OD₆₀₀. Otherwise, a correction has to be performed and this can be done in the software BasyLiCA.

When biological replicates in LCA data are available, two sample algorithms to identify DE genes such as GP2S algorithm (Stegle et al., 2009) or Hotelling T² test (Tai & Speed, 2006) can be applied and the results can be compared with that from the current one-sample method.

After reconstructing biological pathways from time-course gene expression data using reverse engineering techniques, the chosen network inference techniques can be tested for accuracy in edge/interaction calling using non-exposed wild type and mutants; perturbed pathways for different compounds, and concentrations can be inferred and contrasted against their canonical counterparts.

To derive toxicity thresholds based on concentration-pathway alteration relationships using an attractor-based modeling method (Choi et al, 2012), knockout mutants can be used to verify chemical challenge results; pathway-based toxicity thresholds can be compared with thresholds derived from cell growth.

If we can demonstrate a proof-of-concept that pathway alteration is a reliable toxicity endpoint as sensitive as, or more sensitive, than traditional apical or biochemical endpoints, the natural next step would be to demonstrate in vitro and in vivo the causal relationship (dose-dependent pathway perturbation) in more sophisticated eukaryotes ranging from nematodes to zebrafish, mouse, and ultimately humans. These results would allow risk assessors to better incorporate mode of action information, and enable mechanistic toxicology to play a bigger role in next generation risk assessment.

Prior knowledge is available before we reconstruct GRNs of E.coli, as TFs of E.coli are listed in RegulonDB. TFs are usually hub genes because they act as regulators. If prior knowledge can be incorporated into BLOM, the accuracy can be greatly improved.

REFERENCES

- Abeel, T., Helleputte, T., Van de Peer, Y., Dupont, P., & Saeys, Y. (2010). Robust biomarker identification for cancer diagnosis with ensemble feature selection methods. *Bioinformatics*, 26(3), 392-398.
- Aichaoui, L., Jules, M., Le, C. L., Aymerich, S., Fromion, V., & Goelzer, A. (2012). BasyLiCA: a tool for automatic processing of a Bacterial Live Cell Array. *Bioinformatics*, 28(20), 2705-2706.
- Altman, T., Travers, M., Kothari, A., Caspi, R., & Karp, P. D. (2013). A systematic comparison of the MetaCyc and KEGG pathway databases. *BMC Bioinformatics*, 14(1), 112.
- Antonov, A. V. (2011). BioProfiling.de: analytical web portal for high-throughput cell biology. *Nucleic Acids Research*, 39, W323-327.
- Ben-Israel, O., Ben-Israel, H., & Ulitzur, S. (1998). Identification and quantification of toxic chemicals by use of Escherichia coli carrying lux genes fused to stress promoters. *Applied Environmental Microbiology*, 64(11), 4346-4352.
- Bhattacharya, S., Zhang, Q., Carmichael, P. L., Boekelheide, K., & Andersen, M. E. (2011). Toxicity testing in the 21 century: defining new risk assessment approaches based on perturbation of intracellular toxicity pathways. *PLoS One*, 6(6), e20887.
- Brazhnik, P., de la Fuente, A., & Mendes, P. (2002). Gene networks: how to put the function in genomics. *Trends in Biotechnology*, 20(11), 467-472.

- Brohée, S., Faust, K., Lima-Mendez, G., Sand, O., Vanderstocken, G., Deville, Y., & van Helden, J. (2008). NeAT: a toolbox for the analysis of biological networks, clusters, classes and pathways. *Nucleic Acids Research*, 36(suppl 2), W444-W451.
- Cesareni, G., Chatr-aryamontri, A., Licata, L., & Ceol, A. (2008). Searching the MINT database for protein interaction information. *Current Protocol in Bioinformatics*, 8-5.
- Choi, M., Shi, J., Jung, S. H., Chen, X., & Cho, K. H. (2012). Attractor landscape analysis reveals feedback loops in the p53 network that control the cellular response to DNA damage. *Science Signaling*, 5(251), ra83.
- Collins, F. S., Gray, G. M., & Bucher, J. R. (2008). Toxicology. Transforming environmental health protection. *Science*, 319(5865), 906-907.
- Cote, I., Anastas, P. T., Birnbaum, L. S., Clark, R. M., Dix, D. J., Edwards, S. W. et al. (2012). Advancing the next generation of health risk assessment. *Environmental Health Perspectives*, 120(11), 1499-1502.
- Currie, R. A. (2012). Toxicogenomics: the challenges and opportunities to identify biomarkers, signatures and thresholds to support mode-of-action. *Mutation Research*, 746(2), 97-103.
- de Jong, H. (2002). Modeling and simulation of genetic regulatory systems: a literature review. *Journal of Computational Biology*, 9(1), 67-103.
- Doderer, M. S., Yoon, K., & Robbins, K. A. (2010). SIDEKICK: Genomic data driven analysis and decision-making framework. *BMC Bioinformatics*, 11, 611.

- Dorigo, M. (1994). Learning by probabilistic Boolean networks. In *Neural Networks, 1994. IEEE World Congress on Computational Intelligence, 1994 IEEE International Conference on* (Vol. 2, pp. 887-891). IEEE.
- Edwards, S. W., & Preston, R. J. (2008). Systems biology and mode of action based risk assessment. *Toxicological Sciences*, 106(2), 312-318.
- Ehrenreich, A. (2006). DNA microarray technology for the microbiologist: an overview. *Applied Microbiology and Biotechnology*, 73(2), 255-273.
- Elad, T., Lee, J. H., Belkin, S., & Gu, M. B. (2008). Microbial whole-cell arrays. *Microbial Biotechnology*, 1(2), 137-148.
- Elliott, B., Kirac, M., Cakmak, A., Yavas, G., Mayes, S., Cheng, E., . . . Meral Ozsoyoglu, Z. (2008). PathCase: pathways database system. *Bioinformatics*, 24(21), 2526-2533.
- Gambi, N., Pasteris, A., Fabbri, E. (2007). Acetylcholinesterase activity in the earthworm *Eisenia andrei* at different conditions of carbaryl exposure. *Comparative Biochemistry and Physiology Part C: Toxicology and Pharmacology*, 145(4), 678-85.
- Gille, C., Hubner, K., Hoppe, A., & Holzhutter, H. G. (2011). Metannogen: annotation of biological reaction networks. *Bioinformatics*, 27(19), 2763-2764.
- Goffard, N., & Weiller, G. (2007). PathExpress: a web-based tool to identify relevant pathways in gene expression data. *Nucleic Acids Research*, 35(Web Server issue), W176-181.
- Gong, P., Inouye, L. S., & Perkins, E. J. (2007). Comparative neurotoxicity of two energetic compounds, hexanitrohexaazaisowurtzitane and hexahydro-1,3,5-

- trinitro-1,3,5-triazine, in the earthworm *Eisenia fetida*. *Environmental Toxicology and Chemistry*, 26(5), 954-9.
- Gong, P., Pirooznia, M., Guan, X., & Perkins, E. J. (2010). Design, validation and annotation of transcriptome-wide oligonucleotide probes for the oligochaete annelid *Eisenia fetida*. *PLoS One*, 5(12), e14266.
- Gou, N., & Gu, A. Z. (2011). A new Transcriptional Effect Level Index (TELI) for toxicogenomics-based toxicity assessment. *Environmental Science and Technology*, 45(12), 5410-5417.
- Götz, S., García-Gómez, J. M., Terol, J., Williams, T. D., Nagaraj, S. H., Nueda, M. J., ... & Conesa, A. (2008). High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic Acids Research*, 36(10), 3420-35.
- Hong, F., Breitling, R., McEntee, C. W., Wittner, B. S., Nemhauser, J. L., & Chory, J. (2006). RankProd: a bioconductor package for detecting differentially expressed genes in meta-analysis. *Bioinformatics*, 22(22), 2825-2827.
- Huang da, W., Sherman, B. T., Tan, Q., Kir, J., Liu, D., Bryant, D., . . . Lempicki, R. A. (2007). DAVID Bioinformatics Resources: expanded annotation database and novel algorithms to better extract biology from large gene lists. *Nucleic Acids Research*, 35, W169-175.
- Huang, G. T., Athanassiou, C., & Benos, P. V. (2011). mirConnX: condition-specific mRNA-microRNA network integrator. *Nucleic Acids Research*, 39, W416-423.
- Huang, L. T. (2009). An integrated method for cancer classification and rule extraction from microarray data. *Journal of Biomedical Sciences*, 16, 25.

- Jaiswal, P., Ni, J., Yap, I., Ware, D., Spooner, W., Youens-Clark, K., . . . McCouch, S. (2006). Gramene: a bird's eye view of cereal genomes. *Nucleic Acids Research*, 34(Database issue), D717-723.
- Jett, D. A. (2012). Chemical toxins that cause seizures. *Neurotoxicology*, 33(6), 1473-5.
- JJ, G. (2008). Ajax: a new approach to web applications. Retrieved from <http://www.adaptivepath.com/publications/essays/archives/000385.php>
- Kaimal, V., Bardes, E. E., Tabar, S. C., Jegga, A. G., & Aronow, B. J. (2010). ToppCluster: a multiple gene list feature analyzer for comparative enrichment clustering and network-based dissection of biological systems. *Nucleic Acids Research*, 38, W96-102.
- Kalaitzis, A. A., & Lawrence, N. D. (2011). A simple approach to ranking differentially expressed gene expression time courses through Gaussian process regression. *BMC Bioinformatics*, 12, 180.
- Kamburov, A., Pentchev, K., Galicka, H., Wierling, C., Lehrach, H., & Herwig, R. (2011). ConsensusPathDB: toward a more complete picture of cell biology. *Nucleic Acids Research*, 39(Database issue), D712-717.
- Karnovsky, A., Weymouth, T., Hull, T., Tarcea, V. G., Scardoni, G., Laudanna, C., . . . Omenn, G. S. (2012). Metscape 2 bioinformatics tool for the analysis and visualization of metabolomics and gene expression data. *Bioinformatics*, 28(3), 373-380.
- Karp, P. D., Paley, S., & Romero, P. (2002). The Pathway Tools software. *Bioinformatics*, 18 Suppl 1, S225-S232.

- Karp, P. D., Riley, M., Paley, S. M., & Pellegrini-Toole, A. (2002). The MetaCyc Database. *Nucleic Acids Research*, 30(1), 59-61.
- Karp, P. D., Riley, M., Saier, M., Paulsen, I. T., Collado-Vides, J., Paley, S. M. et al. (2002). The EcoCyc Database. *Nucleic Acids Research*, 30(1), 56-58.
- Kawaji, H., Severin, J., Lizio, M., Forrest, A. R., van Nimwegen, E., Rehli, M., . . . Daub, C. O. (2011). Update of the FANTOM web resource: from mammalian transcriptional landscape to its dynamic regulation. *Nucleic Acids Research*, 39(Database issue), D856-860.
- Kerrien, S., Aranda, B., Breuza, L., Bridge, A., Broackes-Carter, F., Chen, C., . . . Hermjakob, H. (2012). The IntAct molecular interaction database in 2012. *Nucleic Acids Research*, 40(Database issue), D841-846.
- Klein, J., Leupold, S., Munch, R., Pommerenke, C., Johl, T., Karst, U., . . . Retter, I. (2008). ProdoNet: identification and visualization of prokaryotic gene regulatory and metabolic networks. *Nucleic Acids Research*, 36(Web Server issue), W460-464.
- Kohl, U., Lotspiech, J., & Kaplan, M. A. (1997). Safeguarding digital library contents and users. *D-lib Magazine*, 3(9).
- Kou, C., & Springsteel, F. (1997). The security mechanism in the World Wide Web (WWW) and the Common Gateway Interface (CGI). Example of Central Police University entrance examination system. In *Security Technology, 1997. Proceedings. The Institute of Electrical and Electronics Engineers 31st Annual 1997 International Carnahan Conference on* (pp. 114-119). IEEE.

- Krewski, D., Acosta, D., Jr., Andersen, M., Anderson, H., Bailar, J. C., III, Boekelheide, K. et al. (2010). Toxicity testing in the 21st century: a vision and a strategy. *Journal of Toxicology and Environmental Health Part B: Critical Reviews*, 13(2-4), 51-138.
- Kuhn, M., Szklarczyk, D., Franceschini, A., von Mering, C., Jensen, L. J., & Bork, P. (2012). STITCH 3: zooming in on protein-chemical interactions. *Nucleic Acids Research*, 40(Database issue), D876-880.
- Laule, O., Hirsch-Hoffmann, M., Hruz, T., Gruissem, W., & Zimmermann, P. (2006). Web-based analysis of the mouse transcriptome using Genevestigator. *BMC Bioinformatics*, 7, 311.
- Li, P. (2009). *Inferring Gene Regulatory Networks from Time Series Microarray Data* (Doctoral Dissertation). The University of Southern Mississippi, Hattiesburg, MS.
- Li, Y., Gong, P., Perkins E. J., Zhang, C., & Wang, N. (2011). RefNetBuilder: a platform for construction of integrated reference gene regulatory networks from expressed sequence tags. *BMC Bioinformatics*, 12 Suppl 10, S20.
- Li, Z., Shaw, S. M., Yedwabnick, M. J., & Chan, C. (2006). Using a state-space model with hidden variables to infer transcription factor activities. *Bioinformatics*, 22(6), 747-754.
- Liang, C., Jaiswal, P., Hebbard, C., Avraham, S., Buckler, E. S., Casstevens, T., ... & Stein, L. (2008). Gramene: a growing plant comparative genomics resource. *Nucleic Acids Research*, 36(suppl 1), D947-D953.

- Lopes, C. T., Franz, M., Kazi, F., Donaldson, S. L., Morris, Q., & Bader, G. D. (2010). Cytoscape Web: an interactive web-based network browser. *Bioinformatics*, 26(18), 2347-2348.
- Lopes, C. T. (2010). Cytoscape Web. Retrieved from <http://cytoscapeweb.cytoscape.org/>
- Ludwig, S., Tinwell, H., Schorsch, F., Cavaille, C., Pallardy, M., Rouquie, D. et al. (2011). A molecular and phenotypic integrative approach to identify a no-effect dose level for antiandrogen-induced testicular toxicity. *Toxicological Sciences*, 122(1), 52-63.
- Melamed, S., Elad, T., & Belkin, S. (2012). Microbial sensor cell arrays. *Current Opinions in Biotechnology*, 23(1), 2-8.
- Murali, T., Pacifico, S., Yu, J., Guest, S., Roberts, G. G., 3rd, & Finley, R. L., Jr. (2011). DroID 2011: a comprehensive, integrated resource for protein, transcription factor, RNA and gene interactions for Drosophila. *Nucleic Acids Research*, 39(Database issue), D736-743.
- Mutwil, M., Obro, J., Willats, W. G., & Persson, S. (2008). GeneCAT--novel webtools that combine BLAST and co-expression analyses. *Nucleic Acids Research*, 36(Web Server issue), W320-326.
- National Research Council (U.S.). Committee on Toxicity Testing and Assessment of Environmental Agents. (2007). *Toxicity testing in the 21st century: a vision and a strategy*. Washington, DC: National Academies Press.
- Obayashi, T., & Kinoshita, K. (2011). COXPRESdb: a database to compare gene coexpression in seven model animals. *Nucleic Acids Research*, 39(Database issue), D1016-1022. doi: 10.1093/nar/gkq1147

- Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H., & Kanehisa, M. (1999). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, 27(1), 29-34.
- Paley, S. M., Latendresse, M., & Karp, P. D. (2012). Regulatory network operations in the Pathway Tools software. *BMC Bioinformatics*, 13, 243. doi: 10.1186/1471-2105-13-243
- Pavlopoulos, G. A., Wegener, A. L., & Schneider, R. (2008). A survey of visualization tools for biological network analysis. *BioData Mining*, 1, 12. doi: 10.1186/1756-0381-1-12
- Penfold, C. A., & Wild, D. L. (2011). How to infer gene networks from expression profiles, revisited. *Interface Focus*, 1(6), 857-870. doi: DOI 10.1098/rsfs.2011.0053
- Pursnani, V. (2001). An introduction to Java servlet programming. *Crossroads*, 8(2), 3-7.
- Rangel, C., Angus, J., Ghahramani, Z., Lioumi, M., Sotheran, E., Gaiba, A. et al. (2004). Modeling T-cell activation using gene expression profiling and state-space models. *Bioinformatics*, 20(9), 1361-1372.
- Reimand, J., Tooming, L., Peterson, H., Adler, P., & Vilo, J. (2008). GraphWeb: mining heterogeneous biological networks for gene modules with functional significance. *Nucleic Acids Research*, 36(Web Server issue), W452-459. doi: 10.1093/nar/gkn230
- Robertson, M. (2004). Reactome: clear view of a starry sky. *Drug Discovery Today*, 9(16), 684-685. doi: 10.1016/S1359-6446(04)03217-9

- Sales, G., Coppe, A., Bisognin, A., Biasiolo, M., Bortoluzzi, S., & Romualdi, C. (2010). MAGIA, a web-based tool for miRNA and Genes Integrated Analysis. *Nucleic Acids Research*, 38(Web Server issue), W352-359. doi: 10.1093/nar/gkq423
- Salgado, H., Peralta-Gil, M., Gama-Castro, S., Santos-Zavaleta, A., Muniz-Rascado, L., Garcia-Sotelo, J. S. et al. (2013). RegulonDB v8.0: omics data sets, evolutionary conservation, regulatory phrases, cross-validated gold standards and more. *Nucleic Acids Research*, 41(Database issue), D203-D213.
- Schuster, S., Fell, D. A., & Dandekar, T. (2000). A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks. *Nature Biotechnology*, 18(3), 326-332.
- Shermin, A., & Orgun, M. A. (2009). Proceedings from ACM symposium on Applied Computing 2009: *Using dynamic Bayesian networks to infer gene regulatory networks from expression profiles*.
- Shimada, T., Murayama, N., Yamazaki, H., Tanaka, K., Takenaka, S., Komori, M. et al. (2013). Metabolic Activation of Polycyclic Aromatic Hydrocarbons and Aryl and Heterocyclic Amines by Human Cytochromes P450 2A13 and 2A6. *Chemical Research in Toxicology*, 26(4), 517-528.
- Shionyu, M., Yamaguchi, A., Shinoda, K., Takahashi, K., & Go, M. (2009). AS-ALPS: a database for analyzing the effects of alternative splicing on protein structure, interaction and network in human and mouse. *Nucleic Acids Research*, 37(Database issue), D305-309. doi: 10.1093/nar/gkn869

- Shmulevich, I., Dougherty, E. R., Kim, S., & Zhang, W. (2002). Probabilistic Boolean Networks: a rule-based uncertainty model for gene regulatory networks. *Bioinformatics*, 18(2), 261-274.
- Smoot, M. E., Ono, K., Ruscheinski, J., Wang, P. L., & Ideker, T. (2011). Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics*, 27(3), 431-432.
- Stark, C., Breitkreutz, B. J., Chatr-Aryamontri, A., Boucher, L., Oughtred, R., Livstone, M. S., . . . Tyers, M. (2011). The BioGRID Interaction Database: 2011 update. *Nucleic Acids Research*, 39(Database issue), D698-704. doi: 10.1093/nar/gkq1116
- Stegle, O., Denby, K. J., Cooke, E.J., Wild, D. L., Ghahramani, Z., & Borgwardt, K. M. (2010). A robust Bayesian two-sample test for detecting intervals of differential gene expression in microarray time series. *Journal of Computational Biology*, 17(3), 355-67.
- Stegle, O., Denby, K., McHattie, S., Meade, S., Wild, D. L., Ghahramani, Z., & Borgwardt, K. (2009). Discovering temporal patterns of differential gene expression in microarray time series. In *German Conference on Bioinformatics 2009*, 28-30. Halle, Germany.
- Steuer, R., Kurths, J., Daub, C. O., Weise, J., & Selbig, J. (2002). The mutual information: detecting and evaluating dependencies between variables. *Bioinformatics*, 18 Suppl 2, S231-240.

- Su, G., Zhang, X., Liu, H., Giesy, J. P., Lam, M. H., Lam, P. K. et al. (2012). Toxicogenomic mechanisms of 6-HO-BDE-47, 6-MeO-BDE-47, and BDE-47 in *E. coli*. *Environmental Science and Technology*, 46(2), 1185-1191.
- Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A. et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of USA*, 102(43), 15545-15550.
- Syed, A. S., D'Antonio, M., & Ciccarelli, F. D. (2010). Network of Cancer Genes: a web resource to analyze duplicability, orthology and network properties of cancer genes. *Nucleic Acids Research*, 38(Database issue), D670-675. doi: 10.1093/nar/gkp957
- Tai, Y. C., & Speed, T. P. (2006). A multivariate empirical Bayes statistic for replicated microarray time course data. *Annals of Statistics*, 34(5), 2387-2412. doi: 10.1214/0090536060000000759
- Tan, Y. (2010). *Statistical Methods for Differential Expressions of Genes Detected in Multiple-Condition Experiment of Microarray*. (M.S. thesis) Texas Medical Center Dissertations (via ProQuest). Accession #: AAI1483404.
- Tannenbaum, L. V. (2012). Is NexGen really the next generation of risk assessment? *Integrated Environmental Assessment and Management*, 8(2), 213-214.
- Thomas, R. S., Clewell, H. J., III, Allen, B. C., Yang, L., Healy, E., & Andersen, M. E. (2012). Integrating pathway-based transcriptomic data into quantitative chemical risk assessment: a five chemical case study. *Mutation Research*, 746(2), 135-143.

- Thomas, R. S., Wesselkamper, S. C., Wang, N. C., Zhao, Q. J., Petersen, D. D., Lambert, J. C. et al. (2013). Temporal concordance between apical and transcriptional points of departure for chemical risk assessment. *Toxicological Sciences*, 134(1), 180-194.
- Tokimatsu, T., Sakurai, N., Ohta, H., Nishitani, K., Koyama, T., Umezawa, T., . . . Shibata, D. (2005). The latest version of the web-based plant metabolic pathway viewer, Kazusa Plant Pathway Viewer (Kappa Viewer). *Plant and Cell Physiology*, 46, S145-S145.
- von Mering, C., Jensen, L. J., Kuhn, M., Chaffron, S., Doerks, T., Kruger, B., . . . Bork, P. (2007). STRING 7--recent developments in the integration and prediction of protein interactions. *Nucleic Acids Research*, 35(Database issue), D358-362. doi: 10.1093/nar/gkl825
- Walia, H., Wilson, C., Condamine, P., Liu, X., Ismail, A. M., Zeng, L., ... & Close, T. J. (2005). Comparative transcriptional profiling of two contrasting rice genotypes under salinity stress during the vegetative growth stage. *Plant Physiology*, 139(2), 822-35.
- Wang, L., Xiong, Y., Sun, Y., Fang, Z., Li, L., Ji, H., & Shi, T. (2010). HLungDB: an integrated database of human lung cancer research. *Nucleic Acids Research*, 38(Database issue), D665-669. doi: 10.1093/nar/gkp945
- Warde-Farley, D., Donaldson, S. L., Comes, O., Zuberi, K., Badrawi, R., Chao, P., . . . Morris, Q. (2010). The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Research*, 38(Web Server issue), W214-220. doi: 10.1093/nar/gkq537

- Williams, L. R., Aroniadou-Anderjaska, V., Qashu, F., Finne, H., Pidoplichko, V., Bannon, D. I., & Braga, M. F. (2011). RDX binds to the GABA(A) receptor-convulsant site and blocks GABA(A) receptor-mediated currents in the amygdala: a mechanism for RDX-induced seizures. *Environmental Health Perspectives*, 119(3), 357-63.
- Windram, O., Madhou, P., McHattie, S., Hill, C., Hickman, R., Cooke, E., ... & Denby, K. J. (2012). Arabidopsis defense against Botrytis cinerea: chronology and regulation deciphered by high-resolution temporal transcriptomic analysis. *Plant Cell*, 24(9), 3530-57.
- Wu, X., Li, P., Wang, N., Gong, P., Perkins, E., Deng, Y., & Zhang, C. (2011). State Space Model with hidden variables for reconstruction of gene regulatory networks. *BMC Systems Biology*, 5 Suppl 3, S3.
- Xenarios, I., Salwinski, L., Duan, X. J., Higney, P., Kim, S. M., & Eisenberg, D. (2002). DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Research*, 30(1), 303-305.
- Xiang, Z., Minter, R. M., Bi, X., Woolf, P. J., & He, Y. (2007). miniTUBA: medical inference by network integration of temporal data using Bayesian analysis. *Bioinformatics*, 23(18), 2423-2432. doi: 10.1093/bioinformatics/btm372
- Yang, Y., Maxwell, A. S., Zhang, X., Wang, N., Perkins, E. J., Zhang, C., & Gong, P. (2013). Differential reconstructed gene interaction networks for deriving toxicity threshold in chemical risk assessment. *BMC Bioinformatics*, 14(Suppl 13), S3.
- Yang, Y., Li, S., Maxwell, A. S., Barker, N. D., Peng, Y., Li, Y., . . . Gong, P. (2013). Proceedings from GENSiPS 2013: *Deciphering Chemically-Induced Reversible*

Neurotoxicity By Reconstructing Perturbed Pathways From Time Series

Microarray Gene Expression Data. Houston, TX, USA.

Yilmaz, A., Mejia-Guerra, M. K., Kurz, K., Liang, X., Welch, L., & Grotewold, E.

(2011). AGRIS: the Arabidopsis Gene Regulatory Information Server, an update.

Nucleic Acids Research, 39(Database issue), D1118-1122.

Zaslaver, A., Bren, A., Ronen, M., Itzkovitz, S., Kikoin, I., Shavit, S. et al. (2006). A

comprehensive library of fluorescent transcriptional reporters for Escherichia coli.

Nature Methods, 3(8), 623-628.

Zhang, X., Wiseman, S., Yu, H., Liu, H., Giesy, J. P., & Hecker, M. (2011). Assessing

the toxicity of naphthenic acids using a microbial genome wide live cell reporter

array system. *Environmental Science and Technology*, 45(5), 1984-1991.

Zou, M., & Conzen, S. D. (2005). A new dynamic Bayesian network (DBN) approach for

identifying gene regulatory networks from time course microarray data.

Bioinformatics, 21(1), 71-79.